JVI Accepts, published online ahead of print on 27 November 2013
J. Virol. doi:10.1128/JVI.03031-13

Evidence of pervasive biologically functional secondary-structures within the genomes of eukaryotic single-stranded DNA viruses

Brejnev Muhizi Muhire[1], Michael Golden[1], Ben Murrell[2], Pierre Lefeuvre[3], Jean-Michel Lett[3], Alistair Gray[4], Art YF Poon[5,6], Nobubelo Kwanele Ngandu[7], Yves Semegni[8], Emil Pavlov Tanov[9], Adérito Luis Monjane[1,4], Gordon William Harkins[9], Arvind Varsani [10,11,12], Dionne Natalie Shepherd[4], Darren Patrick Martin[1,#]

[1]Institute of Infectious Diseases and Molecular Medicine, Computational Biology Group, University of Cape Town, Cape Town, South Africa

[2]Department of Medicine, University of California, San Diego, California, USA

[3]CIRAD, UMR PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, Saint-Pierre, La Réunion, France

[4]Department of Molecular and Cell Biology, University of Cape Town, Rondebosch, Cape Town, South Africa

[5]BC Centre for Excellence in HIV/AIDA, Vancouver, Canada

[6]Department of Medicine, University of British Columbia, Vancouver, Canada

[7]Institute of Infectious Diseases and Molecular Medicine, Division of Medical Virology, University of Cape Town, Cape Town, South Africa

[8]Department of Mathematics and Physics, Cape Peninsula University of Technology, Cape Town, South Africa

[9]South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa

[10]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

[11]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand

[12]Electron Microscope Unit, Division of Medical Biochemistry, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, Cape Town, South Africa

Current address: Institute of Infectious Diseases and Molecular Medicine, Computational Biology Group, University of Cape Town, Cape Town, South Africa

Corresponding Author: darrenpatrickmartin@gmail.com

Running title: Pervasive secondary-structures in ssDNA virus genomes

Abstract word count: 242

Text word count: 7520

Abstract

Single-stranded DNA (ssDNA) viruses have genomes that are potentially capable of forming complex secondary-structures through Watson-Crick base-pairing between their constituent nucleotides. A few of the structural elements formed by such base-pairings are, in fact, known to have important functions during the replication of many ssDNA viruses. What is unknown, however, is (i) whether numerous additional ssDNA virus genomic structural elements predicted to exist by computational DNA folding methods actually exist, and (ii) whether those structures that do exist have any biological relevance. We therefore computationally inferred lists of the most evolutionarily conserved structures within a diverse selection of animal- and plant-infecting ssDNA viruses drawn from the families *Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae* and *Geminiviridae*, and analysed these for evidence of natural selection favouring the maintenance of these structures. While we find evidence that is consistent with purifying selection being stronger at nucleotide sites that are predicted to be base-paired than it is at sites predicted to be unpaired, we also find strong associations between sites that are predicted to pair with one another and site pairs that are apparently coevolving in a complementary fashion. Collectively, these results indicate that natural selection actively preserves much of the pervasive secondary-structure that is evident within eukaryote-infecting ssDNA virus genomes and, therefore, that much of this structure is biologically functional. Lastly, we provide examples of various highly conserved but completely uncharacterised structural elements that likely have important functions within some of the ssDNA virus genomes analysed here.

Introduction

Besides encoding structural, regulatory and enzymatic proteins, the nucleotide sequences of viral genomes encode a wide range of regulatory motifs associated with, amongst other things, transcription (1, 2), translation (3), replication (4) and genome packaging (5). Other types of biologically relevant information encoded by many nucleotide sequences, including those of viruses, are the thermodynamically stable secondary and tertiary structures that these sequences form under physiological conditions.

While the capacity of single-stranded nucleic acid molecules to fold into higher order structures is crucial in all living organisms for the correct functioning of transfer RNA, ribosomal RNA, messenger RNA and small regulatory RNA molecules, such structures are also particularly important in the biology of many viruses with single-stranded DNA (ssDNA) and single-stranded RNA (ssRNA) genomes. Such structures can play vital roles during the entire viral reproductive cycle including the initiation of genome replication (6–12), the regulation of gene expression (13), the control of transcription (14), translation (15–17) and gene splicing (18), and the modulation of host anti-viral responses (19–21).

Within viral genomes, biologically important structural elements tend to be highly conserved across even distantly related species (22). In ssRNA viruses, for example, they include the rev response element (RRE) in Human and Simian immunodeficiency viruses (23–25), the *cis*-acting replication elements (CREs) of flaviviruses (26), luteoviruses (27), carmoviruses (11), coronaviruses (28), alphaflexiviruses (29), reoviruses (30), and picornaviruses (31, 32), the internal ribosomal entry site (IRES) elements of flaviviruses

(33, 34), picornaviruses (35), pestiviruses (36) and dicistroviruses (37), and the cap-independent translation elements (CITEs) found in many plant-infecting ssRNA viruses (38, 39).

Similarly, in many ssDNA virus genomes DNA secondary-structural elements have been identified that have crucial biological functions. While in parvoviruses these include structural elements that are essential for genome replication (9, 40, 41) and optimal gene expression (42–46), in geminiviruses, nanoviruses and circoviruses highly conserved stem-loop structures at their replication origins are essential for the initiation and termination of replication (6, 47–49). Besides these few examples, however, it is currently unknown how pervasive biologically important secondary-structural elements are within the genomes of these viruses (50–52).

It is important to stress the distinction between the simple existence within viral genomes of pervasive stable secondary-structures and the biological importance of these structures. Above a certain length even randomly generated single-stranded oligonucleotides will form stable secondary-structures (53), and it is therefore plausible that many essentially functionless secondary-structural elements might exist within ssDNA and ssRNA viral genomes.

It is, however, theoretically possible to computationally determine the functional importance within viral genomes of secondary-structural elements (detected either by computational prediction or by more rigorous laboratory analyses) by simply examining

patterns of evolution that are evident within groups of related genome sequences. Specifically, although biologically functional secondary-structural elements should be evolutionarily conserved across diverse viral lineages, the nucleotide sequences from which these elements are composed should display distinctive signals of natural selection favouring the maintenance of these structures. Whereas in coding regions these signals might include codon usage biases (54, 55), and decreased rates of synonymous substitution (56), throughout the genome these signals could also include high rates of reversion substitution (51, 57) and increased frequencies of complementarily coevolving nucleotide pairs – particularly amongst those nucleotides predicted to be base-paired within secondary-structural elements (58–60).

Accordingly, experimental investigations of individual structural elements within some ssDNA virus genomes have clearly demonstrated the existence of strong natural selection favouring the maintenance of these elements. For example, when mutations were experimentally introduced that disrupted particular base-pairings within a stem-loop structure at the origin of replication of the circovirus, *Porcine circovirus* 1 (PCV-1), the disrupted base-pairings were rapidly restored during replication through a DNA polymerase mediated template switching mechanism (61, 62). Similarly, in the geminivirus, *Maize streak virus* (MSV), mutations that potentially disrupted base-pairings within a complex computationally predicted structural element were found to very predictably revert to the original nucleotide so as to re-stabilise the structural element (51).

Here we examine the biological relevance of pervasive computationally predicted secondary-structures within diverse eukaryote-infecting ssDNA virus genomes. After using a free energy minimisation approach to identify conserved secondary-structural elements within groups of closely related full genome sequences, we applied various tests to determine whether mutational processes differed between structured and unstructured genome regions in ways consistent with the evolutionary conservation of the identified structural elements. While we provide strong evidence of extensive biologically relevant secondary-structures within eukaryotic ssDNA virus genomes, we further identify what are likely some of the most functionally important uncharacterised structural elements within these genomes.

Materials and methods

Dataset preparation

All available circovirus, anellovirus, parvovirus, nanovirus and geminivirus full genome sequences were obtained from GenBank (http://www.ncbi.nlm.nih.gov/pubmed/) between April 2011 and August 2012. Full genome sequences for each of the five families were preliminarily aligned separately using MUSCLE (63) implemented in MEGA5 (64) and subdivided into datasets of sequences sharing at least 75% sequence identity. This was done to ensure reasonable alignment accuracy during subsequent sequence analyses (65) while at the same time, providing enough sequence diversity to enable the accurate characterisation of evolutionary processes acting to maintain predicted secondary-structures.

A set of 23 datasets was obtained, each containing between 21 and 519 full genome sequences. Each of these full genome sequence datasets was realigned using MUSCLE (with default settings) and, where necessary, manually edited. The resulting alignments will hereafter be referred to as "large datasets" (Table 1). This distinction is important because many of the analyses performed could only be carried out on subsets of these datasets. Specifically, from each of the large datasets we first extracted an "intermediate dataset". In all but four cases each contained one representative sequence from each of the 30 most divergent viral sequence lineages in the large datasets. The exceptional cases were the AnelloTTSuV1, AnelloTTV, ParvoHBoV, and ParvoMPV datasets that respectively contained only 21, 22, 21 and 26 sequences, all of which were included in the intermediate dataset. From each of the intermediate datasets we further extracted a "small dataset" containing one representative sequence of each of the 10 most divergent lineages.

Detection of conserved secondary-structural elements within ssDNA virus genomes

Since biologically relevant secondary-structural elements are likely to be at least partially conserved during evolution, we used the computer program Nucleic Acid Secondary Structure Predictor (NASP; (66)), to identify the conserved secondary-structural elements within the set of representative full genomic sequences in each of the small datasets.

NASP takes as input a set of aligned nucleic acid sequences and uses Gibbs free-energy (67) and Boltzmann probability (68) techniques implemented in the hybrid-ss component of the UNAFOLD software package (69) to determine an ensemble of nearly minimum free

energy (MFE) folds for each of the input sequences. It then uses a nucleotide-shuffling based permutation test to statistically determine the sets of conserved structural elements within the folded sequences that contribute most to their over-all thermodynamic stability.

Precisely, NASP produces sets of "pairing matrices" for each of the input sequences in each small dataset which are then compressed into a "consensus base-pairing matrix," called the M matrix, using a weighted sum of the pairing matrices obtained for each of the input sequences (66). The use by NASP of weighted sums in the calculation of its M Matrix is intended to counteract unavoidable sampling biases in sequence datasets so as to ensure that similar structures within very closely related sequences do not make unfair contributions to the "conservation scores" that NASP calculates for the individual structural elements that it identifies.

Importantly, in our study NASP provided a conservation score for each discrete structural element identified within the M-matrices calculated for each of the small datasets, and indicated the subsets of structural elements referred to as "high confidence structure sets" (HCSSs), that accounted for the analysed sequences having significantly lower MFE scores than those expected in randomised sequences with identical base compositions (66). Whereas the conservation scores for the individual structural elements provided an obvious way of ranking these in order of their likely biological relevance, the demarcation of HCSSs provided an objective means of focusing further analyses into the biological relevance of secondary-structures on just the structural elements that are most likely to really occur.

196

197 In our NASP analysis, sequences were folded as either linear (for the three parvovirus
198 small datasets) or circular (for all 20 other small datasets) ssDNA at either 37^{o}C (for
199 animal-infecting circoviruses, anelloviruses and parvoviruses) or 25^{o}C (for plant-infecting
200 geminiviruses and nanoviruses) under 0M magnesium and 1M sodium ionic conditions.
201 The HCSS was identified using 100 nucleotide shuffling permutations with a permutation
202 *p*-value threshold of 0.05. In all subsequent analyses, the only nucleotides considered as
203 being paired within secondary-structures were those occurring within columns of the large
204 and intermediate dataset nucleotide sequence alignments that corresponded with
205 nucleotides identified by NASP as being paired within the HCSS. Whereas these paired
206 nucleotides were referred to as occurring at "paired-sites", all other nucleotides were
207 referred to as occurring at "unpaired-sites".

208

209 **Neutrality tests for elevated negative selection at paired-sites**

210 Structural elements that increase the fitness of virus genomes are expected to be
211 selectively preserved such that selection disfavouring nucleotide substitutions should be
212 stronger at paired-sites than at unpaired-sites. Specifically, paired-sites might display
213 stronger evidence of purifying selection than unpaired-sites in neutrality tests such as
214 those proposed by Tajima (70) and Fu and Li (71). We calculated Tajima's D and Fu and
215 Li's F statistics for the paired- and unpaired-sites in each of the 23 large datasets.

216

217 Since in all 23 of the analysed large datasets there were invariably fewer paired-sites (i.e.
218 those paired within the HCSS) than there were unpaired-sites (i.e. the remainder of sites in

the various datasets) we devised a permutation test involving the random selection of identical numbers of paired and unpaired-sites and the comparison of summary selection statistics between these paired- and unpaired-site datasets.

From each large dataset we produced 100 datasets each consisting of sites (i.e. entire large dataset alignment columns) randomly sampled with replacement from the pool of unpaired-sites. These permutation datasets contained the same numbers of sites as their corresponding paired-site datasets. Tajima's D and Fu and Li's F statistics were then calculated for all of the paired-site and permutation datasets. For each of the 23 datasets, the probability that purifying selection was operating more strongly on paired-sites than on unpaired-sites was calculated as being approximately equivalent to the proportion of times the D and F statistics calculated for the paired-site dataset were lower than those calculated for the 100 permuted datasets.

Codon-based tests of synonymous substitution rates at paired versus unpaired genomic sites.

Biologically important structural elements that occur within protein-coding sequences are expected to display both selection at the codon level which favours the preservation of amino acid sequences (i.e. selection disfavouring non-synonymous substitutions), and selection at the nucleotide level favouring the maintenance of base-pairing interactions within the structural elements. This "double selection" at codons that contain constituent nucleotides which form base-pairs within biologically important secondary-structures

should result in such codons displaying synonymous substitution rates that are lower than those occurring in codons consisting of unpaired nucleotides.

To determine whether codons corresponding to paired genomic sites displayed significantly lower synonymous substitution rates than those occurring at unpaired genomic sites, nucleotide sequences corresponding to known/suspected genes were extracted from each of the 23 intermediate dataset alignments. Within each of the resulting "gene datasets" all sites encoding amino acids in two or more different frames were removed. Following this, 43 gene datasets containing 200 or more nucleotide sites were retained for further analysis (see Table S1 for details of these datasets). Gene datasets excluded from this set because they retained too few sites included the *ren, mp* and *trap* genes of the GeminiTYLCV, GeminiEACMV and GeminiMYVYV datasets; ORF2 and ORF3 of the AnelloTTSuV1, AnelloTTSuV2 and AnelloTTV datasets; the *vp*1 and *vp2* of the ParvoHBoV dataset; and the *vp1* of the ParvoMPV.

Two methods were used to estimate synonymous substitution rates at individual codon sites within the 43 gene datasets: Partitioning Approach for Inference of Selection (PARRIS) (72) and Fast Unconstrained Bayesian Approximation (FUBAR) (73). Both of these methods apply the time-reversible MG94 codon substitution model which utilises a 61 X 61 codon substitution matrix (74) and both allow independent distributions for synonymous and non-synonymous rates. PARRIS is a random effects likelihood (REL) method permitting the use of only three discrete categories for each rate. FUBAR on the other hand is an approximate Bayesian method which permits the use of many more rate

264 classes (20 in our case) so as to increase the resolution with which, for example, subtle
265 differences in selection pressures operating on individual codons can be distinguished.
266
267 Both FUBAR and PARRIS rely on the use of phylogenetic trees to describe the
268 evolutionary relationships of the sequences being analysed. While it is well established
269 that genetic recombination undermines the accuracy of phylogenetic inference (and by
270 extension many phylogenetics-oriented codon-based selection analysis methods) (75), it
271 was likely that many of the sequences being analysed here were recombinant (76–80). It
272 was therefore necessary to take steps to explicitly account for recombination within these
273 analyses. Accordingly, prior to selection analyses the Genetic Algorithm for Recombination
274 Detection (GARD (81)) method was used to detect recombination breakpoint sites. These
275 sites were then used to partition the alignment into “mostly recombination free” (it is
276 unlikely that every recombination event was detected and accounted for) sub-alignments.
277 For each of these sub-alignments a phylogenetic tree was inferred and these trees were
278 collectively used as inputs for the PARRIS and FUBAR analyses – both of which allow
279 phylogenetic tree topologies and branch lengths to vary across different alignment
280 partitions so as to facilitate accurate inference of natural selection in the presence of
281 recombination (72, 73).
282
283 Within each gene dataset, each codon was categorised as being a “paired-codon” if its
284 third position nucleotide was paired within the relevant HCSS, and an “unpaired-codon” if
285 its third position nucleotide was not a paired nucleotide within the relevant HCSS. Using a
286 Mann-Whitney U test, we determined whether in each of the 43 gene datasets paired-

287 codons had significantly lower synonymous substitution rates than unpaired-codons. All p-
288 values thus calculated were step-down corrected to account for multiple testing.
289
290 **Testing whether paired-sites complementarily coevolve**
291 Mutations at paired-sites may be tolerable within biologically important structural elements
292 if they are followed by compensatory mutations that restore base-pairing (51). We
293 detected evidence of complementary coevolution between pairs of sites within the large
294 datasets using a customised version of the SPIDERMONKEY coevolution script written in
295 HYPHY (82). For any chosen pair of sites within a large dataset, the script compares the
296 standard independent sites 4 X 4 HKY85 nucleotide substitution model (83) to a 16 X 16
297 Muse-Modified HKY85 coevolution model (called M95; (84)) to determine which of these
298 best describes the evolution of individual site pairs. In our case entries in the M95 16 X 16
299 substitution matrix representing changes that potentially maintain base-pairing (including
300 both Watson-Crick pairings such as A-T and G-C and the wobble pair T-G) are multiplied
301 by a pairing factor, λ, and those involving changes between paired- and unpaired-states
302 are multiplied by 1/λ. A maximum likelihood ratio test enabled us to determine whether
303 nucleotide pairs were coevolving. Whereas $\lambda > 1$ for particular coevolving site pairs
304 indicated that they displayed a tendency towards complementary coevolution, $\lambda = 1$
305 indicated a tendency towards site pairs evolving independently and $\lambda < 1$ indicated a
306 tendency towards them coevolving non-complementarily. We identified site pairs
307 displaying strong evidence of complementary coevolution as those with both associated
308 maximum likelihood estimates of $\lambda > 1$, and Muse 95 versus HKY85 likelihood ratio test *p*-
309 values <0.05.

Importantly, due to computational intensity considerations, we restricted our analyses to testing for coevolution only between (i) pairs of sites that were within 100 nucleotides of one another and (ii) pairs of nucleotides that were polymorphic in the input dataset. Also, since recombination can undermine the accuracy with which the phylogenetic trees used to detect coevolution reflect the actual evolutionary relationships of the analysed sequences, we took steps to account for recombination in our analyses. Each large dataset was analysed for recombination using Recombination Detection Program (RDP) version 4.13 (85) which produced a "distributed alignment" in which fragments of recombinant sequences derived from different parental viruses were split up into different sequences. For each of the 23 distributed alignments obtained, a 100 nucleotide sliding window was moved 1 nucleotide step at time across the alignment. At every step the N longest nucleotide sequences were selected (where N is the number of sequences in the original alignment), and saved to an alignment file. Sequences in consecutive windows containing exactly the same N longest sequences were merged into one file (ensuring that no sites from the original alignment were duplicated in the merged alignment). Maximum-likelihood (ML) phylogenetic trees were inferred for each of the resulting alignments under the HKY85 nucleotide substitution model using PhyML3.0 (86). Each of these alignments and their corresponding phylogenetic tree was used as inputs for our complementary coevolution analysis.

Customised computational tools

All computer scripts used in all the analyses conducted can be downloaded from http://web.cbio.uct.ac.za/~brejnev/downloads/Scripts.zip and a customised computer program and datasets for visualisation of predicted structural elements can be downloaded from http://web.cbio.uct.ac.za/~brejnev/downloads/DOOSS.zip (unzip these files and please see the README file for instructions).

Results and Discussion

Numerous evolutionarily conserved secondary-structures are evident within eukaryotic ssDNA virus genomes

We assembled 23 full-genome sequence datasets representing the families *Circoviridae*, *Anelloviridae*, *Parvoviridae*, *Nanoviridae* and *Geminiviridae* (Table 1). In each of these between 69 and 316 conserved secondary-structural elements were identified using the minimum free energy (MFE) approach implemented in the computer program NASP (66). From these lists of conserved structural elements, NASP identified subsets of between five and 132 “high confidence” structural elements. These lists, hereafter referred to as “high confidence structure set” (HCSS) lists, contained those structures primarily responsible for the analysed genomes having greater degrees of predicted structural stability than those of randomised sequences with identical nucleotide compositions. Notably, most of the previously described biologically important structures in these viral genomes were present within the top 30% of structures in the HCSS lists of their respective datasets. These included hairpin structures at the virion strand origins of replication in circoviruses, nanoviruses, and geminiviruses and T-shaped structures required for replication in

parvoviruses. The genomic coordinates of structures within all 23 of the HCSSs were mapped onto their respective genomes (Fig. 1; Fig. 2; Table S2).

Clearly our computational approach for predicting secondary-structure suggests that there exist more conserved genomic secondary-structures within many of these ssDNA virus genomes than is currently appreciated. It is plausible that, as is the case with currently known secondary-structures within these genomes, many of the uncharacterised conserved structures may have been preserved during evolution due to their biological importance.

Although directly testing the biological relevance of any one of the identified potential structures would require detailed mutational analyses of their constituent nucleotides within the context of infectious cloned genomes or analysis of recombinant viral constructs, followed by extensive quantitative fitness assays (87, 88), there are less cumbersome computational approaches for testing whether the identified structures collectively (as opposed to individually) are likely to have any biological relevance. In this regard the biological relevance of the structures in our HCSSs could be tested by comparing how their constituent nucleotides evolve relative to those at positions located outside of the HCSSs.

Therefore, in our subsequent analyses we focused on testing whether, relative to the remainder of nucleotides in the genomes (hereafter referred to as unpaired nucleotides), nucleotides predicted to be base-paired within the HCSS (hereafter referred to as paired

377 nucleotides), are evolving in ways suggestive of their parental structural elements
378 possessing some biological function.
379
380 **Purifying selection is apparently strongest at paired nucleotide sites**
381 Nucleotide sites involved in biologically important base-pairing interactions might be
382 expected to evolve under a greater degree of purifying selection (selection against
383 change) than sites that are not base-paired. Also, sequences evolving under purifying
384 selection are expected to have lower frequencies of minor allele polymorphisms than
385 those evolving under neutral selection and are, therefore, expected to yield negative
386 values for Tajima's D and Fu and Li's F statistics (70, 71). If purifying selection was
387 stronger at base-paired-sites than at unpaired-sites we would expect to see lower values
388 of the D and F statistics for datasets containing only base-paired-sites (constructed from
389 the large dataset alignments by removing all unpaired nucleotide sites) than for datasets
390 containing only unpaired-sites (constructed from the large dataset alignments by removing
391 all paired nucleotide sites).
392
393 In all but one of the 23 large datasets (the exception being the circovirus dataset,
394 CircoDGCV; Table 1), both the paired and unpaired-site alignments consistently yielded
395 negative D and F test static values (see Table 2). In 16/23 of the datasets both the D and
396 F statistics were lower for the paired-site than for the unpaired-site datasets. In 5/23
397 datasets (the anellovirus datasets AnelloTTSuV2 and AnelloTTV, the parvovirus datasets,
398 ParvoAAV, ParvoHBoV and the geminivirus dataset GeminiTYLCV) either the D or F
399 statistics were lower for the paired-site datasets than for unpaired-site datasets. In only

2/23 cases (CircoDGCV and AnelloTTSuV1) did the unpaired-site dataset yield both values of D and F statistics lower than those yielded by the paired-site datasets.

This observation is consistent with our hypothesis that, if paired-sites within the 23 HCSS lists really do reside within biologically important secondary-structures, they should display higher degrees of purifying selection than other sites within the analysed genomes.

However, to test whether values of these statistics were significantly lower at paired-sites than at unpaired-sites in the 21/23 large datasets displaying the expected trend, for each dataset we applied a permutation test involving resampling of identical numbers of sites from the unpaired-site dataset as were present within the paired-site dataset (in each dataset unpaired-sites were invariably more numerous the paired-sites). In each case a *p*-value was computed as the proportion of the 100 permutated unpaired-site datasets that yielded lower D or F values than the corresponding paired-site dataset. In this test a *p*-value <0.05 indicates that you would expect to see a D or F value for an unpaired-site dataset that was lower than that of its corresponding paired-site dataset less than 5% of the time if the null model of neutral evolution was true.

In 11/23 datasets both the D and F statistic permutation tests yielded evidence that paired-sites within the HCSS lists experience significantly stronger (*p*-values < 0.05) purifying selection than the remainder of genomic sites. In a further 6/23 cases, either the D, or F statistic test yielded at least marginal evidence (*p*-values < 0.08) of paired-sites experiencing stronger purifying selection than unpaired-sites. Therefore, in only 6/23

423 cases, was there absolutely no evidence of paired-sites experiencing significantly stronger
424 purifying selection than unpaired-sites.
425
426 Interestingly, all three of the analysed anellovirus datasets were among the six datasets
427 with no evidence of purifying selection acting on paired-sites. It is perhaps also noteworthy
428 that of the eleven datasets displaying strong evidence of base-pairing associated with
429 purifying selection, only two (both of them circoviruses, CircoCoCV and CircoBFDV) were
430 from the ten mammal- and bird-infecting virus datasets. While it is not possible to directly
431 compare the plant- and animal-infecting virus datasets to one another, it is plausible that
432 increased structural stability afforded by the lower physiological temperatures of plants
433 relative to animals might contribute to the genomic structures of the plant viruses being
434 more evolutionarily stable than those of their warm-blooded animal counterparts. A more
435 mundane explanation, however, could simply be that our animal virus datasets were, in
436 general, substantially smaller than our plant virus datasets and that our analysis therefore,
437 simply lacked sufficient power to differentiate between the numbers of low frequency
438 polymorphisms within the paired and unpaired dataset fractions.
439
440 Regardless of possible differences between animal and plant viruses, collectively these
441 results indicate that a substantial proportion of paired-sites within at least 17/23 of the
442 HCSSs are evolving in a manner that is consistent with many of these structures being
443 evolutionarily preserved.
444
445 **Synonymous substitution rates are unusually low at paired genomic sites**

We hypothesised that selection favouring the maintenance of base-pairing within secondary-structures might be particularly evident when these structures occurred within protein-coding regions of the genome. Essentially, we investigated whether codons in which third codon position nucleotides were predicted to be base-paired within the HCSSs had significantly lower synonymous substitution rates than those with unpaired nucleotides in the third codon position.

Synonymous substitution rates at individual codon sites within 43 gene datasets (Table S1) were inferred using the random effects likelihood selection analysis methods, PARRIS (75) and FUBAR (73). These methods indicated that in 27/43 of these datasets, the median substitution rates of codons with paired third position nucleotides were significantly lower than those of codons with unpaired third position nucleotides (multiple comparison corrected Mann-Witney U-test p-value < 0.05). An additional five datasets yielded similar evidence but only with one of the two selection analysis methods (Table 3).

The results of these analyses therefore strongly support our hypothesis that two layers of selection – one operating at the amino acid sequence level and the other at the nucleotide sequence level – are likely acting on nucleotide sites within the HCSSs that fall within coding regions. This suggests not only that many of the predicted secondary-structures represented within the HCSSs really do exist (either within single-stranded genomic DNAs themselves, or within the RNA transcripts that are produced from them), but that these structures likely also make a substantial contribution to the fitness of the genomes within which they reside.

While evidence of lower degrees of nucleotide polymorphism and decreased synonymous substitution rates at paired-sites than at unpaired-sites provides strong support for the existence of many of the predicted secondary-structural elements within the HCSSs, it must be stressed that this result does not necessarily imply that these elements are biologically functional. The reason for this is that besides influencing which arising mutations are deleterious and which are neutral (and, therefore, which mutations are likely to be purged from populations by natural selection), the presence of secondary-structures within ssDNA genomes could potentially also influence the basal rates at which sites within these genomes become mutated (89), simply because base-paired nucleotides might be predisposed to lower mutation rates than their unpaired counterparts (90, 91).

In short term evolution experiments mutations tend to preferentially accumulate at unpaired-sites

If paired-sites within the HCSSs really do form base-pairs within genomic secondary-structures, we hypothesized that these sites might accumulate fewer mutations than unpaired-sites. We tested this hypothesis using mutation data from a series of previously published short-term evolution experiments. In one experiment infectious cloned genomes of two *Maize streak virus* isolates (called MSV-MatA and MSV-VW) closely related to those in the GeminiMSV dataset, were used to infect maize plants (92). In another experiment infectious cloned genomes of a *Tomato yellow leaf curl virus* isolate (called TYX) and a *Tomato leaf curl Comoros virus* isolate (called TOX; both closely related to sequences included in the GeminiTYLCV dataset) were used to infect tomato plants (52).

While over 101 days post-infection the MSV-MatA and MSV-VW genomes were noted to have accumulated 41 and 33 mutations, respectively, at 52 distinct nucleotide sites, over 120 days the TYX and TOX genomes had respectively accumulated 31 and 105 mutations at 135 distinct nucleotide sites. As described previously for our small datasets, we predicted the secondary-structures of each genome pair using NASP in order to obtain, for each pair, its own specific HCSS. We used these HCSSs to construct two-by-two contingency tables for paired-sites (sites predicted to be paired within the HCSS) and unpaired-sites (all sites in the genome other than the HCSS paired-sites) versus variable sites (those where mutations occurred) and invariable sites (those where mutations did not occur) and used these in a Fisher's exact test (93), to assess whether variable sites were significantly clustered outside rather than inside paired-sites.

For MSV-MatA and MSV-VW 11/52 variable sites (~21%) were located at paired nucleotide sites (939/2641 or ~36% of considered sites) within the HCSS, yielding significant evidence (*p*-value = 0.019) that mutations tended to occur more frequently at unpaired nucleotides. Similarly, for TYX and TOX only 5/135 variable sites (~4%) were located at paired nucleotide sites (237/2724 or ~9% of considered sites) within the HCSS regions, indicating a significant tendency (*p*-value = 0.021) for mutations to accumulate more frequently at unpaired nucleotide sites.

Although no analogous experimental data is currently available for any of the other plant- and animal-infecting ssDNA viruses investigated here, it is nevertheless important that even in short term geminivirus evolution experiments such as these, where selection has

514 not had prolonged periods to purge slightly deleterious mutations, there remains such an
515 obvious trend for mutations to preferentially occur at unpaired-sites.
516
517 Unfortunately, even though these experiments were short-term (lasting between 101 and
518 120 days), it remains possible that selection, in addition to a decreased biochemical
519 predisposition to mutation, was responsible for the relatively lower mutation frequencies at
520 paired-sites within these genomes. While still consistent with our hypothesis that selection
521 is acting on secondary-structures to maintain their biological functionality, these results
522 suggest that the alternative hypothesis – that base-paired-sites within secondary-
523 structures are simply biochemically predisposed to mutate more slowly than unpaired-sites
524 – is also entirely plausible.
525
526 Therefore, although we had established up to this point that secondary-structures are
527 likely quite pervasive within ssDNA virus genomes, we were unable to definitively attribute
528 the apparent evolutionary conservation of these structures to natural selection favouring
529 the maintenance of their biological functionality.
530
531 **Base-paired-sites tend to complementarily coevolve**
532 It is expected that, independent of different basal mutation rates at paired- and unpaired-
533 sites, nucleotide substitutions that occur at paired-sites within biologically functional
534 secondary-structures might only be tolerable if coupled with complementary substitutions
535 that reconstitute base-pairing. Therefore, in order to test for natural selection acting to
536 maintain secondary-structures without the confounding effects of base-pairing dependent

basal mutation rate variation, we directly tested for evidence of paired-sites within the HCSSs coevolving with one another in a manner consistent with the maintenance of their base-pairing. Specifically, we tested for associations between sites predicted to be base-paired within the HCSSs and sites detectably coevolving in a complementary fashion within the 23 large datasets. For each large dataset we performed a two-by-two contingency test of site pairs predicted to be paired versus unpaired on the one hand, and sites predicted to be coevolving versus not coevolving on the other.

In all but one circovirus dataset, CircoCoCV, we found strong significant associations (multiple testing corrected *p*-values <0.0001) between paired-sites within the HCSSs and sites for which complementary coevolution was detected (Table 4). It is noteworthy that the CircoCoCV was one of the two animal-infecting virus datasets displaying both strong evidence of base-pairing associated negative selection, and evidence of strong selection disfavouring synonymous substitutions at paired codon sites within coding regions. Therefore, the lack of significant evidence of coevolution between nucleotides predicted to be paired within the CircoCoCV HCSS may simply be due to strong selection disfavouring any substitutions at these sites.

Besides providing additional evidence that many of the structures represented within the HCSSs really do form either within the genomes of these ssDNA viruses, or within their RNA transcripts, this result provides the most compelling evidence yet that natural selection is favouring the maintenance of a substantial proportion of these structures. The simple fact that many of the structures represented within the HCSSs likely provide

significant fitness advantages to the genomes in which they occur, in turn, suggests that many of these structures have as yet, undetermined biological functions.

Potentially important structural elements within eukaryotic ssDNA virus genomes

Whereas we provided evidence of pervasive evolutionarily conserved (and therefore, likely biologically functional) secondary-structures within the various ssDNA virus genomes that we have analysed, we have not up to this point examined any of the individual computationally inferred structural elements in any significant detail. Fortunately, some of the analyses that we performed provide a straightforward means of ranking the identified structures within the HCSSs in order of their likely biological functionality (94). Specifically, these rankings were based on: (1) the degree to which structural elements were conserved across the analysed genomes; (2) the degree to which synonymous substitution rates were constrained at codon sites containing nucleotides that are predicted to be base-paired and; (3) the degree to which nucleotides predicted to be base-paired coevolve with one another. Rankings based collectively on these three criteria are hereafter, referred to as “consensus rankings” (Table S2).

The ten highest ranked structural elements based on these criteria within each of the 23 analysed HCSSs are plotted in magenta and cyan in Fig. 1 and Fig. 2, and are listed in Table S2. It is important to point out that although these top ranked structures contributed most to the signals detected in our earlier association tests, it is possible that many of them do not actually exist in the exact form that we have inferred either in the ssDNA genomes themselves, or in the RNA molecules transcribed from these genomes. Besides

583 expected inaccuracies in the computational inference of DNA and RNA secondary-
584 structures (95), it is likely that even if these structural elements have been accurately
585 inferred, the exact base-pairing configurations within the presented consensus structures
586 will likely vary between the different genomes within each of the analysed datasets. Also, it
587 is very likely that, even within an individual genome, many of these structures will not be
588 static but will instead represent a single reasonably stable base-pairing configuration
589 amongst a (potentially very large) ensemble of similarly stable alternative configurations. It
590 should therefore, be borne in mind that the actual base-pairing interactions within the
591 tertiary structures represented by many of these structural elements, might vary as the
592 structural elements continually transition between their alternative forms.
593
594 Among the individual structural elements that achieved the highest consensus rankings
595 were all of the well-characterised secondary-structures found at the origins of replication of
596 circoviruses (ranks 1 to 6), nanoviruses (ranks 8 to 28), geminiviruses (ranks 1 to 12) and
597 parvoviruses (ranks 1 to 35) (Table S2).
598
599 Additional well-characterised structures detected include the replication associated protein
600 gene (*rep*) intron associated structure (GeminiMSV; rank 16) (51), the parvovirus
601 transcription attenuation stem-loop structures (ParvoMPV; ranks 17 and 34; Table S2)
602 (42), the 3' complementary strand T-shaped structure that binds to the viral capsid in some
603 parvoviruses (ParvoMPV; rank 3; Table S2) (96).
604

605 Besides these well-known structures, we sought to identify other uncharacterised, but
606 likely biologically functional, structural elements within some of these genomes. Rather
607 than exhaustively enumerating every predicted secondary-structural element that might
608 have some biological relevance, we instead focus here on a few examples of the elements
609 that have apparently been conserved across multiple, highly divergent viral lineages in the
610 various viral families that we analysed.
611
612 Geminivirus
613 We identified a particularly conserved 126 to 157 nt secondary-structure within the
614 movement protein (*mp)* gene of all five analysed mastrevirus datasets (GeminiMSV,
615 GeminiPanSV, GeminiWDV, GeminiTYDV-CpCV and GeminiCpCDV; structure G1 in Fig.
616 1 and Fig. 3). In all of these datasets other than GeminiMSV, the entire structure was
617 within the HCSS (7th out of 25 in GeminiCpCDV, 10th out of 47 in GeminiTYDV-CpCV, 21st
618 out of 31 in GeminiPanSV, 16th out of 30 in GeminiWDV, and 51st in GeminiMSV). The
619 structure in the GeminiMSV dataset displayed a particularly high degree of conformational
620 similarity with that in the GeminiPanSV dataset with the two structures sharing a nearly
621 identical 21-nucleotide long stem sequence (Fig. 3) indicating that they are almost
622 certainly homologous. Although the sequences within this structure differ substantially
623 between the other mastrevirus datasets, they all contain the splice donor, acceptor and
624 branch sites previously identified (or predicted) in mastrevirus *mp* introns (97) (Fig. 3),
625 suggesting that the structure is possibly functional within the *mp* mRNA transcript where it
626 might facilitate *mp* intron splicing. Also, likely acceptor and donor sites identified within
627 these various sequences tend to occur at junctions between paired and unpaired

628 nucleotides – a factor which might enhance the accessibility of these sites during splicing
629 (18, 98, 99).
630
631 Another highly conserved secondary-structure that is most likely functional within
632 geminivirus genomes was identified near the 3' end of the coat protein (*cp*) genes of
633 begomoviruses in the GeminiTYLCV, GeminiEACMV and GeminiMYVYV datasets
634 (structure G2 in Fig. 1 and Fig. 4). This structure contains a conserved stem-loop
635 sequence immediately 3' of the *cp* stop codon that contains the likely polyadenylation
636 signals of both virion and complementary strand RNA transcripts (Fig. 4). It is likely
637 therefore, that this structure may be functional either within ssDNA as a transcriptional
638 terminator, or within transcribed mRNA during polyadenylation.
639
640 Parvovirus
641 We identified a variety of uncharacterised parvovirus genomic and/or mRNA structural
642 elements with potential functionality at the start of the large non-structural *(ns1)* gene
643 (Structure P1 in Fig. 2 and Fig. 5; 20th out of 70 HCSS structures in ParvoAAV, 30th out of
644 105 HCSS structures in ParvoHBoV and 34th out of 132 HCSS structures in ParvoMPV),
645 the start of the major virion/viral protein (*vp1*) gene (Structure P2 in Fig. 2 and Fig. 5; 16th
646 out of 70 HCSS structures in ParvoAAV and 9th out of 132 HCSS structures in ParvoMPV),
647 the start of the small non-structural (*np1)* gene (structure P3 in Fig. 2 and Fig. 5, 59th out of
648 105 HCSS structure in ParvoHBoV) and the start of the minor virion protein (*vp2*) gene
649 (structure P4 in Fig. 2 and Fig. 5; 56th out of 132 HCSS structures in ParvoMPV). Although
650 there were no sequence similarities shared between positionally analogous structures in

651 the different parvovirus datasets, this was not unexpected given that these datasets
652 represent species within different genera (with sequences in different datasets sharing on
653 average only 57.8% sequence identity). The ParvoMPV IR-*ns1* structure contains a stem-
654 loop identified to play role in transcription attenuation of *Parvovirus minute virus of mice*
655 (42) (structure P1; Fig. 5). In this regard, it is noteworthy that start codons within the
656 structures that we have identified are consistently located either within, or immediately
657 adjacent to unpaired loop or bulge regions (Fig. 5). This tendency was also noted in other
658 datasets analysed, and it is plausible that structures spanning the start codons of genes in
659 these different families are functional within either partially single stranded DNA during the
660 initiation of transcription, or in transcribed mRNA during the initiation of translation.
661
662 Circovirus
663 While we were unable to identify any secondary-structures that were clearly conserved
664 across all five circovirus datasets analysed, within the moderately divergent CircoCoCV
665 and CircoBFDV datasets (these two datasets share on average 64% pairwise sequence
666 identity), we identified an intergenic region (IR) stem-loop structure (structure C1 in Fig. 2
667 and Fig. 6), which is highly conserved in each of the respective datasets (ranked 5th out of
668 35 HCSS structures in CircoCoCV and 1st out 41 HCSS structures in CircoBFDV). Despite
669 the sequences of this structure sharing no obvious similarity between the two datasets, in
670 both datasets the stem is GC rich (and therefore, predicted to be very stable) with a loop
671 sequence containing a conserved pentanucleotide (CGAAG). This structural element
672 could potentially contain the complementary strand replication origin or it might be

functional either during the termination of transcription, or in the post-transcriptional processing of mRNA transcripts.

Anellovirus

A conserved T-shaped structure was identified in the IRs of the two anellovirus datasets, AnelloTTSuV1 and AnelloTTuSV2 (structure A1 in Fig. 2 and Fig. 7; 7th out of 20 structures in the AnelloTTSuV1 HCSS and 1st out of 27 structures in the AnelloTTSuV2 HCSS). Even though these two datasets are moderately divergent (sequences within them share on average 60.7% pairwise identity), the structure is strikingly conserved between the two datasets. In both datasets it has a nearly identical predicted T-shaped conformation with a highly conserved 17-nucleotide long sequence at the top of the "T" (highlighted in sky blue in Fig. 7).

Given the high degree to which this structure has been conserved between these two moderately divergent anellovirus datasets, we attempted to identify a homologous structure within our third highly divergent anellovirus dataset (AnelloTTV). The most likely homologue of this structure also resides within the IR and is ranked 7th out of 78 structures in the AnelloTTV HCSS (structure A1 in Fig. 2 and Fig. 7). However, the AnelloTTV structure has a stem-loop rather than a T-shaped configuration and lacks the 17-nucleotide sequence that is conserved in the AnelloTTSuV1 and AnelloTTSuV2 structures. All three Anellovirus structures nevertheless, contain two similar sequences (five- and six nucleotide long) at similar positions within their stems (outlined in blue and yellow in Fig. 7), which strongly suggests that these structures are indeed homologous.

696

697 Unlike with many other circular ssDNA viruses that replicate by rolling circle replication, it
698 is presently unknown where the Anellovirus virion and complementary strand origins of
699 replication (*oris*) reside. Given that the virion strand *oris* of other ssDNA viruses generally
700 occur within IRs and have a characteristic stem-loop structure with an A-T rich loop
701 sequence, it is plausible that this highly conserved Anellovirus structural element might
702 contain the Anellovirus virion strand *ori*. However, characterisation of replication
703 competent sub-full-length *Torque teno virus* (TTV) genomes (which are closely related to
704 those represented in our AnelloTTV dataset) has suggested that the TTV virion strand *ori*
705 is approximately 470 nucleotides 3' of the highly conserved TTV stem-loop structure that
706 we have identified here (in the region of a small stem-loop structure ranked 83rd, bellow
707 the HCSS in our AnelloTTV dataset; data not shown) (100). Importantly, the structure we
708 have identified falls outside the genomic region that is conserved within these defective
709 genomes and, in the TTV genome at least, is therefore, unlikely to be the virion strand *ori*.
710 Apart from possibly containing the virion strand *ori*, this highly conserved structural
711 element could alternatively be involved in either complementary strand replication or
712 transcriptional regulation, both of which are also carried out by IR sequences in all other
713 known ssDNA viruses.

714

715 **Conclusion**

716 Using computational methods we have identified numerous secondary-structures that
717 probably form at least transiently, within eukaryotic ssDNA virus genomes, and shown that
718 a significant proportion of these predicted structures are likely biologically functional (Table

719 5). We have further provided a few examples of currently uncharacterised genomic
720 secondary-structures which, due to high degrees of evolutionary conservation across
721 multiple highly divergent viral lineages, likely play a central role in the biology of the
722 various ssDNA viruses examined here.
723
724 Although we found evidence consistent with natural selection strongly disfavouring the
725 accumulation of substitutions at paired-sites, we also found that paired-sites tended to
726 display lower nucleotide variability than unpaired-sites. Using data from published
727 evolution experiments, we showed that, in at least one of the analysed virus families (the
728 geminiviruses), it is possible that this discrepancy may simply be due to mutation
729 frequencies at paired-sites being lower than those at unpaired-sites (possibly due to base-
730 paired nucleotides being less mutable than unpaired nucleotides). We were nevertheless
731 able to clearly demonstrate the action of selection by showing that those base-paired-sites
732 which do accumulate mutations display a significant tendency towards complementary
733 coevolution with their predicted pairing partners – presumably to maintain the biological
734 function of their parent structures.
735
736 Despite providing compelling evidence of pervasive biologically functional secondary-
737 structures within eukaryote-infecting ssDNA viruses, it is important to reiterate that our
738 study has certain limitations. It is very likely that the complex genomic structures of these
739 viruses are not entirely static. The secondary and tertiary structures of these entire
740 genomes are, in fact, very likely to shift continually between large numbers of different
741 thermodynamically stable states. We cannot therefore, be absolutely certain if the

742 computationally predicted structures identified here are a good reflection of those which
743 form most commonly within these ssDNA virus genomes. Also, although examples of
744 individual genomic structural elements that are highly conserved across divergent virus
745 lineages are likely to have some biological functionality, we cannot know without further
746 laboratory experimentation either what the precise functions of these structures might be,
747 or whether they function within the context of ssDNA or transcribed RNA.
748
749 Regardless of whether specific individual structures form, or are functional within ssDNA
750 or transcribed RNA molecules, it is absolutely clear from our study that, at the whole-
751 genome scale, selection favouring the overall maintenance of pervasive biologically
752 functional nucleic acid secondary-structures has likely been a major theme in the
753 evolutionary history of eukaryotic ssDNA viruses.
754

Acknowledgments

BMM is funded by the University of Cape Town, South Africa. PL and JML are funded by the Conseil Régional de La Réunion, European Union (FEDER) and Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD). ALM is funded by the National Research Foundation (South Africa) and the Carnegie Corporation of New York. BM is funded by the CFAR Translational Virology Core: P30 AI036214, Molecular Epidemiology (Avant Garde grant): DP1 DA034978. AFYP is supported by a New Investigator Award from the Canadian Institutes of Health Research (Canadian HIV Vaccine Initiative) and by a Scholar Award from the Michael Smith Foundation for Health Research / St. Paul's Hospital Foundation - Providence Health Care Research Institute. DNS is funded by the Pannar (Pty) Ltd. DPM and GWH are funded by the South African National Research Foundation. The authors would like to thank the Centre for High Performance Computing in Cape Town and the Information Communication Technology Services Department at the University of Cape Town for use of their high-performance computing clusters.

References

1. **Yuen L, Moss B**. 1987. Oligonucleotide sequence signaling transcriptional termination of vaccinia virus early genes. Proc. Natl. Acad. Sci. U. S. A. **84**:6417–21.

2. **Hefferon KL, Moon Y-S, Fan Y**. 2006. Multi-tasking of nonstructural gene products is required for bean yellow dwarf geminivirus transcriptional regulation. FEBS J. **273**:4482–94.

3. **Shen R, Miller WA**. 2004. The 3' untranslated region of tobacco necrosis virus RNA contains a barley yellow dwarf virus-like cap-independent translation element. J. Virol. **78**:4655–64.

4. **Song SI, Miller WA**. 2004. cis and trans Requirements for Rolling Circle Replication of a Satellite RNA. J. Virol. **78**:3072–3082.

5. **Stockley PG, Twarock R, Bakker SE, Barker AM, Borodavka A, Dykeman E, Ford RJ, Pearson AR, Phillips SE V, Ranson N a, Tuma R**. 2013. Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. J. Biol. Phys. **39**:277–87.

6. **Steinfeldt T, Finsterbusch T, Mankertz A**. 2001. Rep and Rep' protein of porcine circovirus type 1 bind to the origin of replication in vitro. Virology **291**:152–60.

7. **Berns KI**. 1990. Parvovirus replication. Microbiol. Rev. **54**:316–29.

8. **Gronenborn B**. 2004. Nanoviruses: genome organisation and protein function. Vet. Microbiol. **98**:103–109.

9. **Ashktorab H, Srivastava A**. 1989. Identification of nuclear proteins that specifically interact with adeno-associated virus type 2 inverted terminal repeat hairpin DNA. J. Virol. **63**:3034–9.

10. **Faurez F, Dory D, Grasland B, Jestin A**. 2009. Replication of porcine circoviruses. Virol. J. **6**:60.

11. **Sun X, Simon AE**. 2006. A cis-replication element functions in both orientations to enhance replication of Turnip crinkle virus. Virology **352**:39–51.

12. **Mohan BR, Dinesh-Kumar SP, Miller WA**. 1995. Genes and cis-acting sequences involved in replication of barley yellow dwarf virus-PAV RNA. Virology **212**:186–95.

13. **Ilyinskii PO, Schmidt T, Lukashev D, Meriin AB, Thoidis G, Frishman D, Shneider AM**. 2009. Importance of mRNA secondary structural elements for the expression of influenza virus genes. OMICS **13**:421–30.

14. **Koev G, Mohan BR, Miller WA**. 1999. Primary and secondary structural elements required for synthesis of barley yellow dwarf virus subgenomic RNA1. J. Virol. **73**:2876–85.

15. **Guo L, Allen EM, Miller WA**. 2001. Base-pairing between untranslated regions facilitates translation of uncapped, nonpolyadenylated viral RNA. Mol. Cell **7**:1103–9.

16. **Zuo X, Wang J, Yu P, Eyler D, Xu H, Starich MR, Tiede DM, Simon AE**. 2009. Solution structure of the cap-independent translational enhancer and ribosome-binding element in the 3 0 UTR of turnip crinkle virus. Biophys. Comput. Biol. **107**:1385–1390.

17. **Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM**. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature **460**:711–6.

18. **Moss WN, Dela-Moss LI, Priore SF, Turner DH**. 2012. The influenza A segment 7 mRNA 3' splice site pseudoknot/hairpin family. RNA Biol. **9**:1305–10.

19. **Wikström FH, Meehan BM, Berg M, Timmusk S, Elving J, Fuxler L, Magnusson M, Allan GM, McNeilly F, Fossum C**. 2007. Structure-dependent modulation of alpha interferon production by porcine circovirus 2 oligodeoxyribonucleotide and CpG DNAs in porcine peripheral blood mononuclear cells. J. Virol. **81**:4919–27.

20. **Wikström FH, Fossum C, Fuxler L, Kruse R, Lövgren T**. 2011. Cytokine induction by immunostimulatory DNA in porcine PBMC is impaired by a hairpin forming sequence motif from the genome of Porcine Circovirus type 2 (PCV2). Vet. Immunol. Immunopathol. **139**:156–66.

21. **Simmonds P, Tuplin A, Evans DJ**. 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA **10**:1337–51.

22. **Simon AE, Gehrke L**. 2009. RNA conformational changes in the life cycles of RNA viruses, viroids, and virus-associated RNAs. Biochim. Biophys. Acta **1789**:571–83.

23. **Fernandes J, Jayaraman B, Frankel A**. 2012. The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. RNA Biol. **9**:6–11.

24. **Powell DM, Amaral MC, Wu JY, Maniatis T, Greene WC**. 1997. HIV Rev-dependent binding of SF2/ASF to the Rev response element: possible role in Rev-mediated inhibition of HIV RNA splicing. Proc. Natl. Acad. Sci. U. S. A. **94**:973–8.

25. **Le SY, Malim MH, Cullen BR, Maizel J V**. 1990. A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. Nucleic Acids Res. **18**:1613–23.

26. **You S, Stump DD, Branch AD, Rice CM**. 2004. A cis-Acting Replication Element in the Sequence Encoding the NS5B RNA-Dependent RNA Polymerase Is Required for Hepatitis C Virus RNA Replication. J. Virol. **78**:1352–1366.

27. **Koev G, Liu S, Beckett R, Miller WA**. 2002. The 3'-Terminal Structure Required for Replication of Barley Yellow Dwarf Virus RNA Contains an Embedded 3' End. Virology **292**:114–126.

28. **Raman S, Bouma P, Williams GD, Brian DA**. 2003. Stem-loop III in the 5' untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. J. Virol. **77**:6720–30.

29. **Pillai-Nair N, Kim K-H, Hemenway C**. 2003. Cis-acting Regulatory Elements in the Potato Virus X 3′ Non-translated Region Differentially Affect Minus-strand and Plus-strand RNA Accumulation. J. Mol. Biol. **326**:701–720.

30. **Chen D, Barros M, Spencer E, Patton JT**. 2001. Features of the 3'-consensus sequence of rotavirus mRNAs critical to minus strand synthesis. Virology **282**:221–9.

31. **Paul A V, Rieder E, Kim DW, van Boom JH, Wimmer E**. 2000. Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg. J. Virol. **74**:10359–70.

32. **Nagashima S, Sasaki J, Taniguchi K**. 2003. Functional analysis of the stem-loop structures at the 5′ end of the Aichi virus genome. Virology **313**:56–65.

33. **Piñeiro D, Martinez-Salas E**. 2012. RNA structural elements of hepatitis C virus controlling viral RNA translation and the implications for viral pathogenesis. Viruses **4**:2233–50.

34. **Thurner C**. 2004. Conserved RNA secondary structures in Flaviviridae genomes. J. Gen. Virol. **85**:1113–1124.

35. **Pelletier J, Sonenberg N**. 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature **334**:320–5.

36. **Kolupaeva VG, Pestova T V, Hellen CU**. 2000. Ribosomal binding to the internal ribosomal entry site of classical swine fever virus. RNA **6**:1791–807.

37. **Kanamori Y, Nakashima N**. 2001. A tertiary structure model of the internal ribosome entry site (IRES) for methionine-independent initiation of translation. RNA **7**:266–74.

38. **Miller WA, Wang Z, Treder K**. 2007. The amazing diversity of cap-independent translation elements in the 3'-untranslated regions of plant viral RNAs. Biochem. Soc. Trans. **35**:1629–33.

39. **Simon AE, Miller WA**. 2013. 3' cap-independent translation enhancers of plant viruses. Annu. Rev. Microbiol. **67**:21–42.

40. **Cossons N, Faust EA, Zannis-Hadjopoulos M**. 1996. DNA polymerase delta-dependent formation of a hairpin structure at the 5' terminal palindrome of the minute virus of mice genome. Virology **216**:258–64.

41. **Sun Y, Chen AY, Cheng F, Guan W, Johnson FB, Qiu J**. 2009. Molecular characterization of infectious clones of the minute virus of canines reveals unique features of bocaviruses. J. Virol. **83**:3956–67.

42. **Perros M, Spegelaere P, Dupont F, Vanacker JM, Rommelaere J**. 1994. Cruciform structure of a DNA motif of parvovirus minute virus of mice (prototype strain) involved in the attenuation of gene expression. J. Gen. Virol. **75 (Pt 10**:2645–53.

43. **Krauskopf A, Bengal E, Aloni Y**. 1991. The block to transcription elongation at the minute virus of mice attenuator is regulated by cellular elongation factors. Mol. Cell. Biol. **11**:3515–21.

44. **Ben-Asher E, Aloni Y**. 1984. Transcription of minute virus of mice, an autonomous parvovirus, may be regulated by attenuation. J. Virol. **52**:266–76.

45. **Resnekov O, Aloni Y**. 1989. RNA polymerase II is capable of pausing and prematurely terminating transcription at a precise location in vivo and in vitro. Proc. Natl. Acad. Sci. U. S. A. **86**:12–6.

46. **Bohenzky RA, LeFebvre RB, Berns KI**. 1988. Sequence and symmetry requirements within the internal palindromic sequences of the adeno-associated virus terminal repeat. Virology **166**:316–27.

47. **Orozco BM, Hanley-Bowdoin L**. 1996. A DNA structure is required for geminivirus replication origin function. J. Virol. **70**:148–58.

48. **Hafner GJ, Stafford MR, Wolter LC, Harding RM, Dale JL**. 1997. Nicking and joining activity of banana bunchy top virus replication protein in vitro. J. Gen. Virol. **78 (Pt 7)**:1795–9.

49. **Cheung AK**. 2006. Rolling-circle replication of an animal circovirus genome in a theta-replicating bacterial plasmid in Escherichia coli. J. Virol. **80**:8686–94.

50. **Morozov Sy, Chernov B, Merits A, Blinov V**. 1994. Computer-assisted predictions of the secondary structure in the plant virus single-stranded DNA genome. J. Biomol. Struct. Dyn. **11**:837–847.

51. **Shepherd DN, Martin DP, Varsani A, Thomson JA, Rybicki EP, Klump HH**. 2006. Restoration of native folding of single-stranded DNA sequences through reverse mutations: an indication of a new epigenetic mechanism. Arch. Biochem. Biophys. **453**:108–22.

52. **Martin DP, Lefeuvre P, Varsani A, Hoareau M, Semegni J, Dijoux B, Vincent C, Reynaud B, Lett J**. 2011. Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS Pathog. **7**:e1002203.

53. **Schultes EA, Spasic A, Mohanty U, Bartel DP**. 2005. Compact and ordered collapse of randomly generated RNA sequences. Nat. Struct. Mol. Biol. **12**:1130–6.

54. **Hasegawa M, Yasunaga T, Miyata T**. 1979. Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res. **7**:2073–9.

55. **Cardinale DJ, DeRosa K, Duffy S**. 2013. Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. Viruses **5**:162–81.

56. **Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D, Seoighe C**. 2008. Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol. J. **5**:160.

57. **Cheung AK**. 2005. Detection of rampant nucleotide reversion at the origin of DNA replication of porcine circovirus type 1. Virology **333**:22–30.

58. **Cheng N, Mao Y, Shi Y, Tao S**. 2012. Coevolution in RNA molecules driven by selective constraints: evidence from 5S rRNA. PLoS One **7**:e44376.

59. **Fernández N, Fernandez-Miragall O, Ramajo J, García-Sacristán A, Bellora N, Eyras E, Briones C, Martínez-Salas E**. 2011. Structural basis for the biological relevance of the invariant apical stem in IRES-mediated translation. Nucleic Acids Res. **39**:8572–85.

60. **Hofacker I, Fekete M, Flamm C, Huyenen M, Rauscher S, Stolorz P, PF S**. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucleic Acids Res. **26**:3825–3836.

61. **Cheung AK**. 2004. Detection of template strand switching during initiation and termination of DNA replication of porcine circovirus. J. Virol. **78**:4268–77.

62. **Cheung AK**. 2004. Palindrome regeneration by template strand-switching mechanism at the origin of DNA replication of porcine circovirus via the rolling-circle melting-pot replication model. J. Virol. **78**:9016–29.

63. **Edgar RC**. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–7.

64. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S**. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**:2731–9.

65. **Wilm A, Mainz I, Steger G**. 2006. An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms Mol. Biol. **1**:19.

66. **Semegni JY, Wamalwa M, Gaujoux R, Harkins GW, Gray A, Martin DP**. 2011. NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics **27**:2443–5.

67. **Greiner W, Neise L, Stöcker H**. 1995. Thermodynamics and statistical mechanics, p. 101. Springer-Verlag, New York. http://www.tandar.cnea.gov.ar/~sacanell/Greiner.pdf.

68. **Ding Y**. 2003. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31**:7280–7301.

69. **Markham NR, Zuker M**. 2008. UNAFold: software for nucleic acid folding and hybridization., p. 3–31. *In* Keith, JM (ed.), Methods in molecular biology (Clifton, N.J.). Humana Press, Totowa, NJ.

70. **Tajima F**. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics **123**:585–595.

71. **Fu YX, Li WH**. 1993. Statistical tests of neutrality of mutations. Genetics **133**:693–709.

72. **Scheffler K, Martin DP, Seoighe C**. 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics **22**:2493–2499.

73. **Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K**. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. **30**:1196–205.

74. **Muse S V, Gaut BS**. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–24.

75. **Scheffler K, Martin DP, Seoighe C**. 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics **22**:2493–9.

76. **Lefeuvre P, Lett J-M, Varsani A, Martin DP**. 2009. Widely conserved recombination patterns among single-stranded DNA viruses. J. Virol. **83**:2697–707.

77. **Julian L, Piasecki T, Chrzastek K, Walters M, Muhire B, Harkins GW, Martin DP, Varsani A**. 2013. Extensive recombination detected among beak and feather disease virus isolates from breeding facilities in Poland. J. Gen. Virol. **94**:1086–95.

78. **Padidam M, Sawyer S, Fauquet CM**. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology **265**:218–25.

79. **Shackelton L a, Hoelzer K, Parrish CR, Holmes EC**. 2007. Comparative analysis reveals frequent recombination in the parvoviruses. J. Gen. Virol. **88**:3294–301.

80. **Navas-Castillo J, Sánchez-Campos S, Noris E, Louro D, Accotto GP, Moriones E**. 2000. Natural recombination between Tomato yellow leaf curl virus-is and Tomato leaf curl virus. J. Gen. Virol. **81**:2797–801.

81. **Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW**. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics **22**:3096–8.

82. **Pond SLK, Frost SDW, Muse S V**. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**:676–9.

83. **Hasegawa M, Kishino H, Yano T**. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:169–174.

84. **Muse S V**. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics **139**:1429–39.

85. **Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P**. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **26**:2462–3.

86. **Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O**. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**:307–21.

87. **McCormack JC, Simon AE**. 2004. Biased hypermutagenesis associated with mutations in an untranslated hairpin of an RNA virus. J. Virol. **78**:7813–7.

88. **Staplin WR, Miller WA**. 2008. In vivo analyses of viral RNA translation. Methods Mol. Biol. **451**:99–112.

89. **Simmonds P, Smith DB**. 1999. Structural constraints on RNA virus evolution. J. Virol. **73**:5787–94.

90. **Frederico L, Kunkel T, Shaw B**. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry **29**:2532–2537.

91. **Xia X, Yuen KY**. 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. BMC Genet. **6**:20.

92. **Monjane AL, Pande D, Lakay F, Shepherd DN, van der Walt E, Lefeuvre P, Lett J-M, Varsani A, Rybicki EP, Martin DP**. 2012. Adaptive evolution by recombination is not associated with increased mutation rates in Maize streak virus. BMC Evol. Biol. **12**:252.

93. **Fisher RA**. 1922. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. J. R. Stat. Soc. **85**:87.

94. **Golden M, Martin D**. 2013. DOOSS: a tool for visual analysis of data overlaid on secondary structures. Bioinformatics **29**:271–2.

95. **Ray SS, Pal SK**. 2013. RNA secondary structure prediction using soft computing. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**:2–17.

96. **Willwand K, Hirt B**. 1991. The minute virus of mice capsid specifically recognizes the 3' hairpin structure of the viral replicative-form DNA: mapping of the binding site by hydroxyl radical footprinting. J. Virol. **65**:4629–35.

97. **Wright E, Heckel T, Groenendijk J, Davies J, Boulton M**. 1997. Splicing features in maize streak virus virion- and complementary-sence gene expression. Plant J. **12**:1285–1297.

98. **Warf MB, Berglund JA**. 2010. Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. **35**:169–78.

99. **Munroe SH**. 1984. Secondary structure of splice sites in adenovirus mRNA precursors. Nucleic Acids Res. **12**:8437–56.

100. **De Villiers E-M, Borkosky SS, Kimmel R, Gunst K, Fei J-W**. 2011. The diversity of torque teno viruses: in vitro replication leads to the formation of additional replication-competent subviral molecules. J. Virol. **85**:7284–95.

Table and Figure legends

Table 1. List of the 23 large datasets analysed.

Table 2. Tajima's D and Fu and Li F statistics for paired and unpaired genomic site alignments.

Table 3. Comparison of synonymous substitution rates at paired and unpaired codon-sites.

Table 4. Tests for associations between paired-sites and complementarily coevolving sites.

Table 5. Summary of the various tests carried on the 23 different ssDNA virus genome sequence datasets.

Fig. 1. Secondary-structure map of plant-infecting ssDNA viruses.

Genome organisation maps of plant-infecting ssDNA viruses. In each map, the arcs represent the coordinates of identified structural elements within "high confidence structure sets" (HCSS). These highly conserved structural elements are those primarily responsible for the estimated structural stability of the analysed genomes being greater than that of randomised sequences with identical nucleotide compositions. The ten structures collectively displaying the greatest degrees of base-pairing conservation, lowest associated synonymous substitution rates and greatest degrees of complementary coevolution between paired nucleotides are shown using arcs in cyan and magenta (to distinguish the two complementary parts of the stem sequences). All remaining structures are shown using black arcs. Black arrows indicate examples of currently uncharacterised but likely biologically functional structures that are apparently conserved across multiple

datasets (coloured in green and brown when these were not ranked among the top ten within their respective HCSS). The known secondary-structural elements at the virion strand origins of replication are indicated by a star symbol at the 12 o'clock position of the genomes. Numbers in brackets indicate the lengths of the genomes in kilobases (kb) and the ratio of the numbers of high confidence structures over the total numbers of predicted secondary-structures. Italicized abbreviations of gene names: *rep*=replication associated protein; *cp* = coat protein; *mp* = movement protein; *clink* = cell cycle link protein; *nsp* = nuclear shuttle protein; *ren* = replication enhancer protein; *trap* = transcription activator protein.

Fig. 2. Secondary-structure map of animal-infecting ssDNA viruses.

Genome organisation maps of animal-infecting ssDNA viruses. In each map, the arcs (for circular genomes) and vertical lines (for linear genomes) represent the coordinates of identified structural elements within "high confidence structure sets" (HCSS). The ten structures collectively displaying the greatest degrees of base-pairing conservation, lowest associated synonymous substitution rates and greatest degrees of complementary coevolution between paired nucleotides are shown using arcs in cyan and magenta (to distinguish the two complementary parts of the stem sequences). All remaining structures are shown using black arcs/vertical lines. Black arrows indicate examples of currently uncharacterised but likely biologically functional structures that are apparently conserved across multiple datasets (coloured in green and brown when these were not ranked among the top ten structures within their respective HCSS). The known secondary-structural elements at the virion strand origins of replication are indicated by a star symbol

1086 at the 12 o'clock position of the genomes. Numbers in brackets indicate the lengths of the
1087 genomes in kilobases (kb) and the ratio of the numbers of high confidence structures over
1088 the total numbers of predicted secondary-structures. Italicized abbreviations represent the
1089 gene names encoding the following proteins: *rep*=replication associated protein; *cp* = coat
1090 protein; *ns1* = large non-structural protein; *np*1 = small non-structural protein; *vp1* = major
1091 virion/viral protein; *vp2* = minor virion/viral protein and *ORF* = unnamed open reading
1092 frame.
1093
1094 **Fig. 3. Secondary-structure associated with the intron of the mastrevirus movement**
1095 **protein gene.**
1096 A secondary-structure associated with the movement protein gene intron was predicted in
1097 all five mastrevirus datasets. This structure is highly conserved and contains splice donor
1098 and acceptor sites (indicated by arrows), as well as, in the case of the GeminiMSV,
1099 GeminiPanSV and GeminiWDV, a likely lariat sequences (outlined in pink). The similarities
1100 between these structures include homologous stem-loop structures conserved in all but
1101 GeminiWDV (highlighted in blue and yellow), a highly conserved stem-structure found in
1102 both GeminiMSV and GeminiPanSV, and conserved sequences in the stems of
1103 GeminiTYDV-CpCV and GeminiCpCDV. The rank ratio shows the actual rank of a
1104 structure over the total number of structures predicted in the high confidence structure set
1105 (HCSS). This structure is highly ranked in GeminiCpCDV and GeminiTYDV-CpCV (ranked
1106 7th out of 25 structures in HCSS and 10th out of 47 structures in the HCSS set
1107 respectively). In case of GeminiCpCDV, base-pairing interactions displaying significant
1108 associated complementary coevolution (*p*-value <0.05) are represented by a red line

where the degree of redness reflects the *p*-value. Whereas nucleotide sequence variability is reflected by a sequence logo at each position, each position is also associated with a colour ranging from blue to green depicting the rate of synonymous substitutions of the codon site at which the nucleotide is located. Low synonymous substitution rates are observable in the stem region in all datasets, indicating that there is a high degree of conservation at these particular sites. Although the sequence of this structure is divergent in all five mastrevirus datasets, it is plausible that this structure has some function during splicing of the movement protein intron.

Fig. 4. Secondary-structure associated with the 3' end of the begomovirus coat protein gene.

A secondary-structure with a potential role in transcriptional termination was predicted at the end of the coat protein gene of the begomovirus datasets, GeminiTYLCV, GeminiEACMV and GeminiMYVYV. In all these, the structure has a stop codon and a stem-loop containing a polyadenylation signal (the complementary polyadenylation signalling sequences within the stem-loops are in bold text). A common stem-loop structure between the GeminiEACMV and GeminiMYVYV dataset is highlighted in yellow. Nucleotide logos and colours respectively indicate degrees of sequence variability and associated synonymous nucleotide substitution rates as outlined in Fig. 3. Nucleotides falling outside genes are shaded grey. Base-pairing interactions displaying significant associated complementary coevolution (*p*-value <0.05) are represented by a red line where the degree of redness reflects the *p*-value. The rank ratio shows the actual rank of a structure over the total number of structures predicted in the high confidence structure set.

Fig. 5. Parvovirus secondary-structures predicted at the start of genes.

Secondary-structures predicted at the start of genes represented in the parvovirus datasets ParvoAAV, ParvoHBoV and ParvoMPV are shown. These include those spanning the start of the large non-structural proteins (*ns1;* P1), the major viral/virion proteins (*vp1;* P2), the small non-structural protein (*np1;* P3) and the minor viral/virion proteins (*vp2,* P4). Nucleotide logos and colours respectively indicate degrees of sequence variability and associated synonymous nucleotide substitution rates as outlined in Fig. 3. Base-pairing interactions displaying significant associated complementary coevolution (*p*-value <0.05) are represented by a line where the degree of redness reflects the p-value. The rank ratio shows the actual rank of a structure over the total number of structures predicted in the high confidence structure set. The ParvoMPV IR-*ns1* stem-loop involved in transcription attenuation is highlighted in grey. In the depicted structures start codons are consistently located either within or immediately adjacent to an unpaired loop or bulge, which might enhance the accessibility of these codons during transcription or translation.

Fig. 6. Conserved circovirus stem-loop structure within the intergenic region.

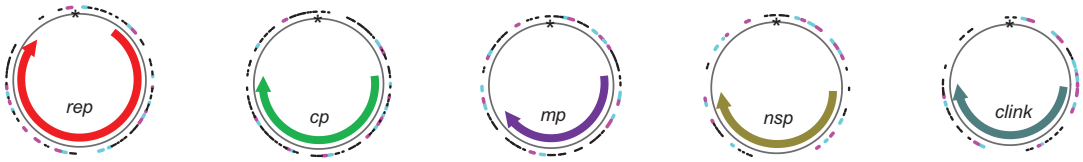
A stem-loop structures that are highly conserved within the intergenic regions of the two circovirus datasets, CircoCoCV and CircoBFDV, is shown. Nucleotide logos reflect degrees of sequence variability. The stems of these structures have high GC-contents and display clear evidence of complementary coevolution between base-paired nucleotides within the CircoBFDV dataset (base-pairing interactions displaying significant associated

complementary coevolution with *p*-value <0.05 are represented by red lines where the degree of redness reflects the p-value). Additionally, a pentanucleotide loop (highlighted in green) is highly conserved at the top of the stem in both datasets.

Fig. 7. Anellovirus highly conserved intergenic T-shaped structures.

A T-shaped structure predicted within the intergenic region (IR) of two divergent anellovirus datasets, AnelloTTSuV1 and AnelloTTSuV2 is shown. These structures have homologous 17-nucleotide long sequences on top of the "T" (highlighted in sky blue) and similar sequences in the stem (highlighted in green, yellow and blue). The homologue to these structures in the even more divergent AnelloTTV dataset has a stem-loop rather than a "T" configuration. It shares similar sequences (highlighted using yellow and blue) with the ones found in the other anellovirus datasets. In the AnelloTTV structure, base-pairing interactions displaying significant associated complementary coevolution (*p*-value <0.05) are represented by a red line where the degree of redness reflects the p-value. Nucleotide logos reflect the degree of sequence diversity at individual sites.

Nanovirus



NanoBBTVR [EF546807] (1.1 kb, 31/76)

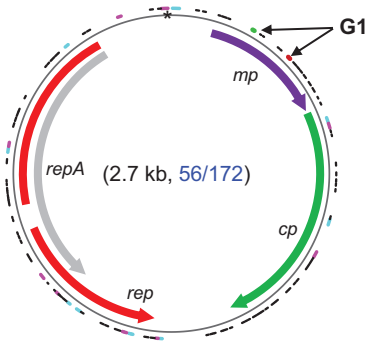
NanoBBTVS [AB113661] (1.1 kb, 49/80)

NanoBBTVM [JF957665] (1.1 kb, 32/71)

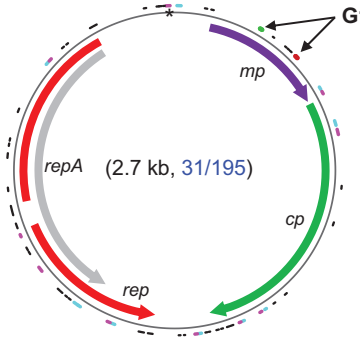
NanoBBTVN [AF238879] (1.1 kb, 19/83)

NanoBBTVC [EF520722] (1 kb, 24/69)

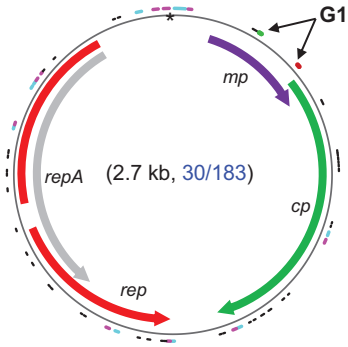
Geminivirus



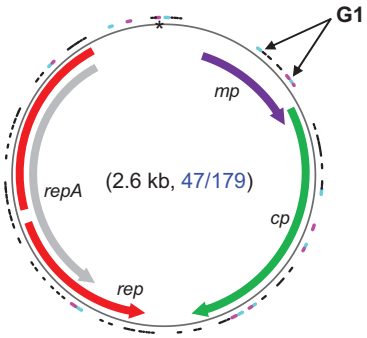
GeminiMSV [Y00514]



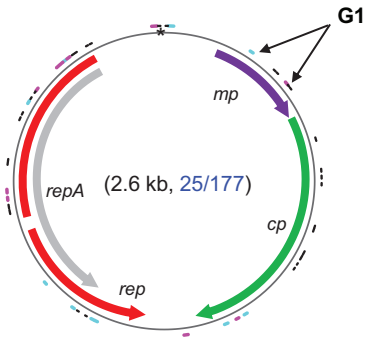
GeminiPanSV [L39638]



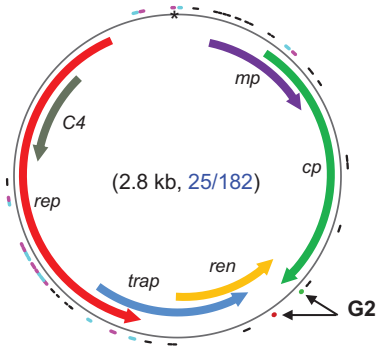
GeminiWDV [FJ620684]



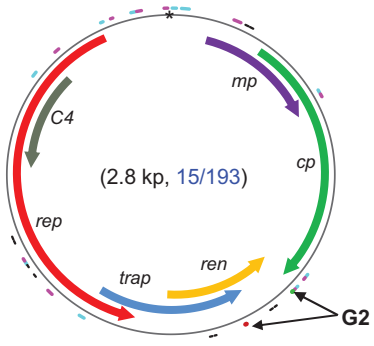
GeminiTYDV-CpCV [JN989413]



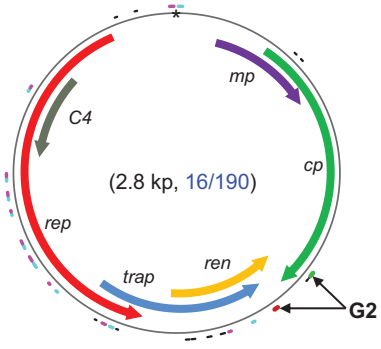
GeminiCpCDV [AM849097]



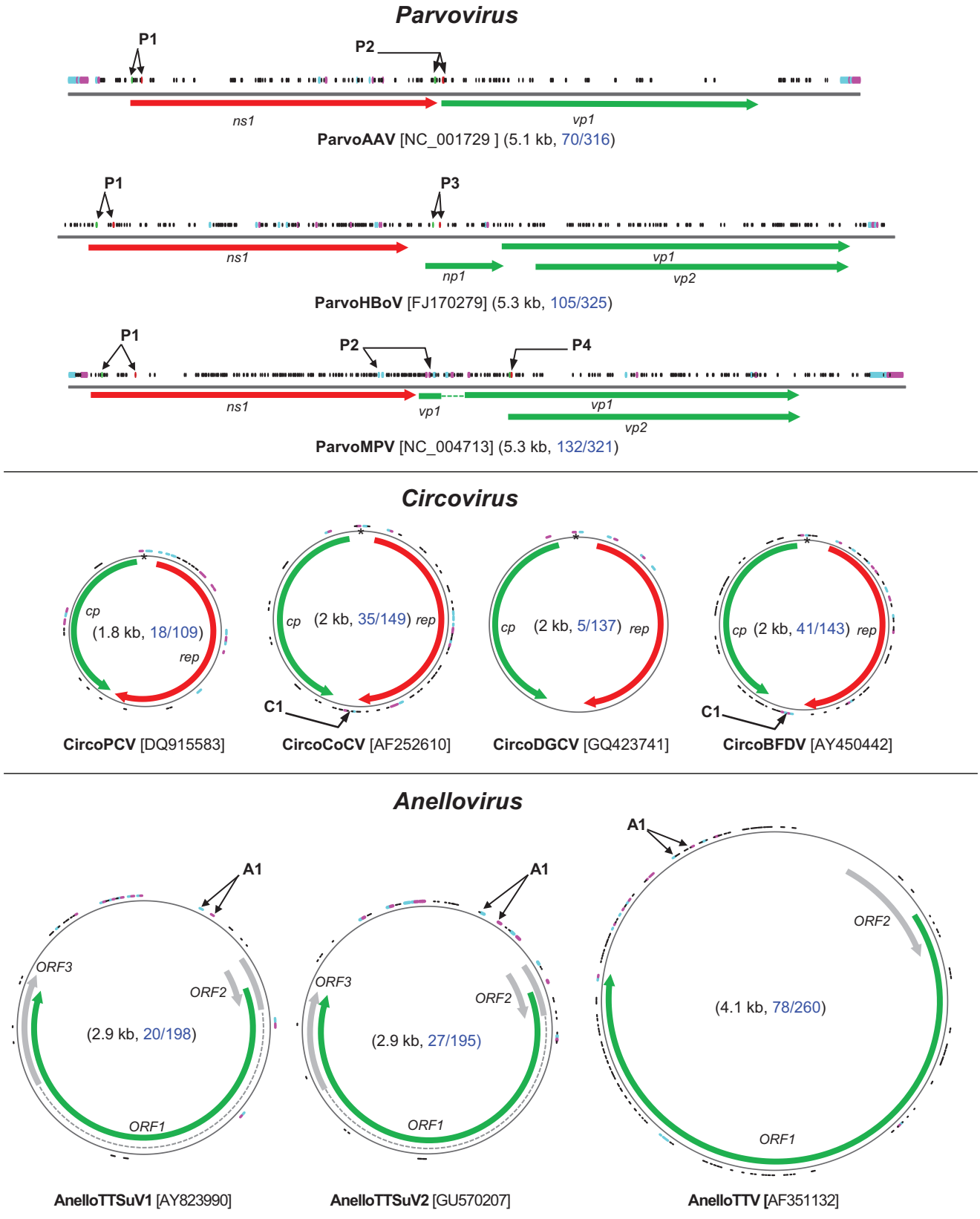
GeminiTYLCV [EU847740]

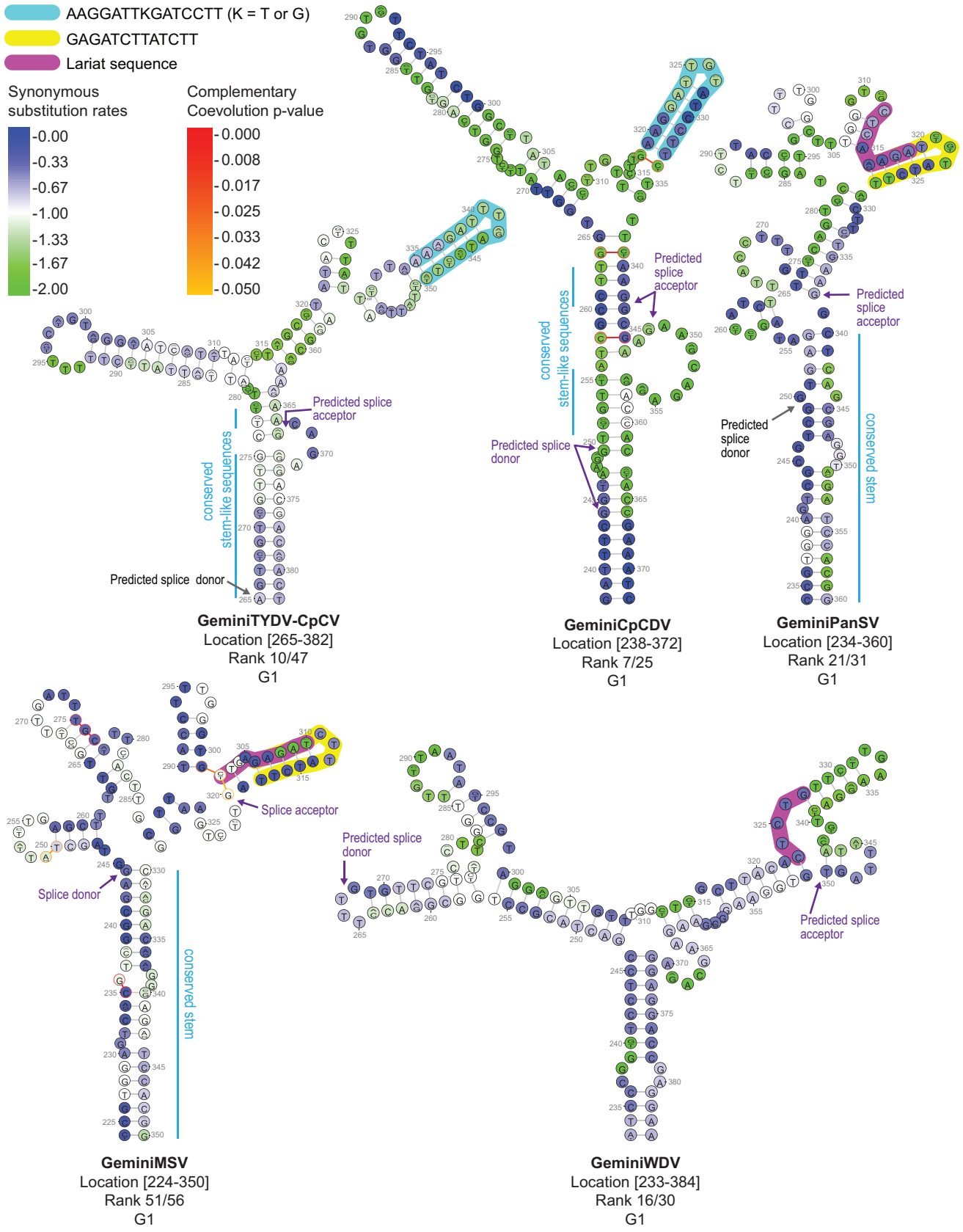


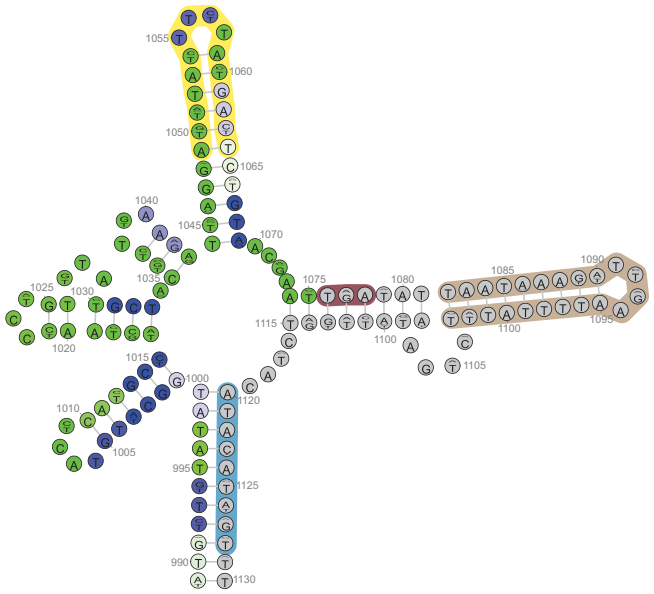
GeminiEACMV [AJ717558]



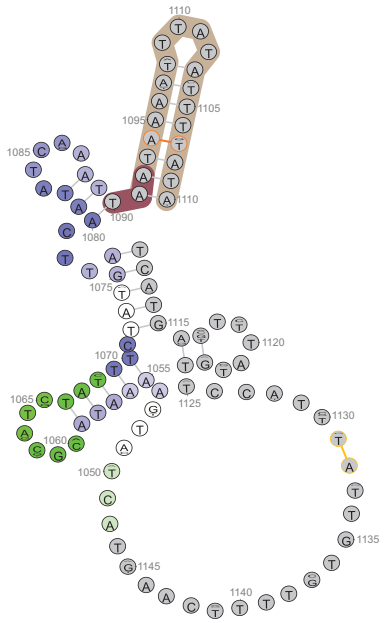
GeminiMYVYV [EF185318]



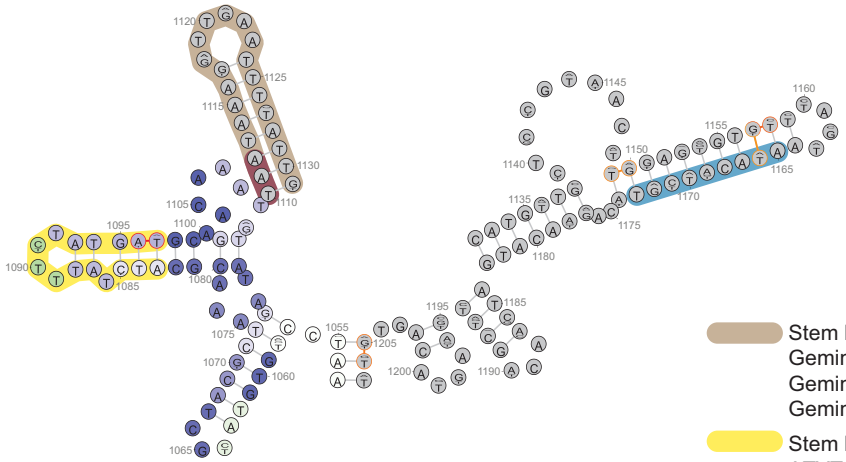




GeminiMYVYV
Location [989-1130]
Rank 13/16
G2



GeminiTYLCV
Location [1048-1146]
rank 21/25
G2



GeminiEACMV
Location [1053-1207]
Rank 14/15
G2

Synonymous substitution rates

-0.00
-0.33
-0.67
-1.00
-1.33
-1.67
-2.00

Complementary Coevolution p-value

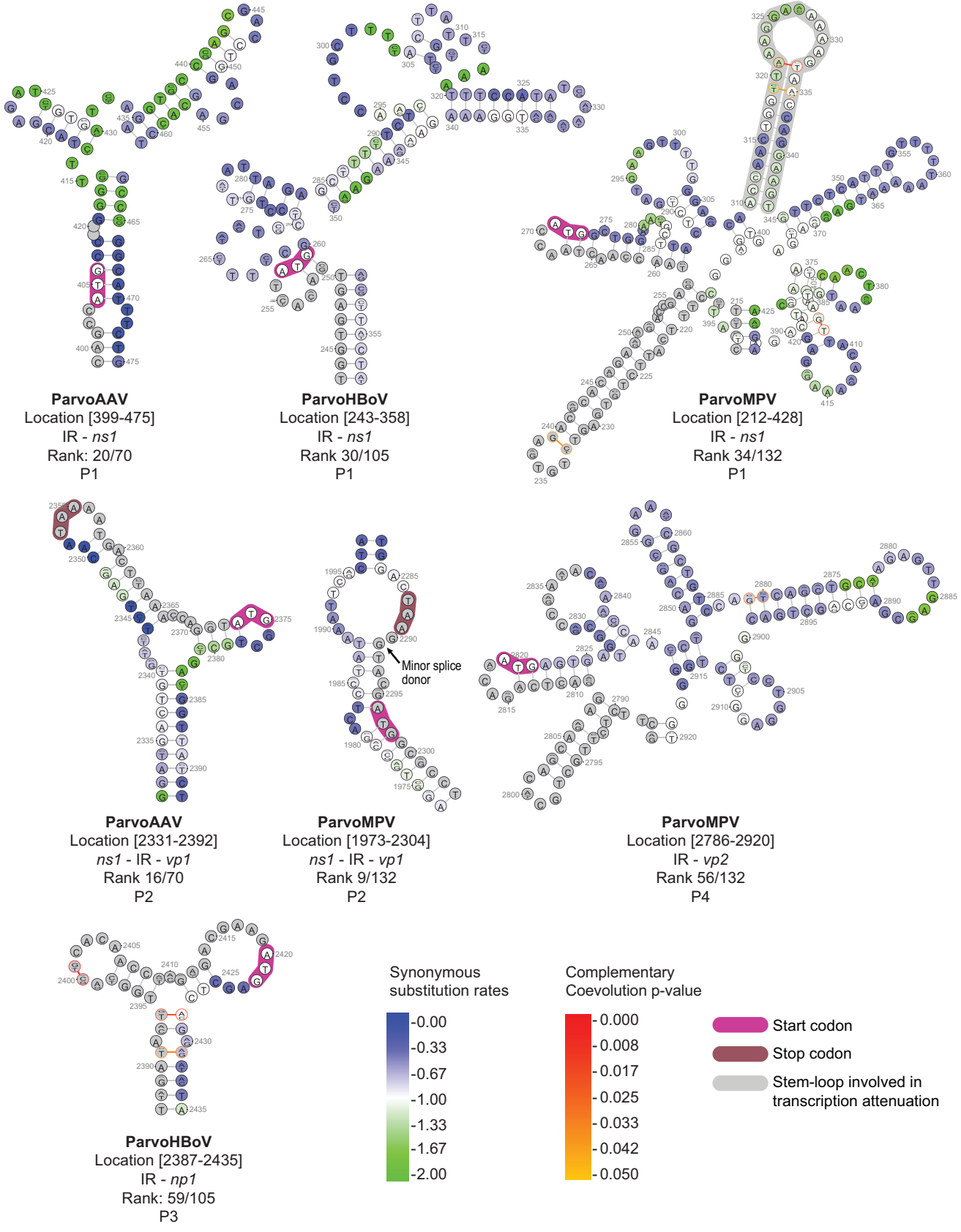
-0.000
-0.008
-0.017
-0.025
-0.033
-0.042
-0.050

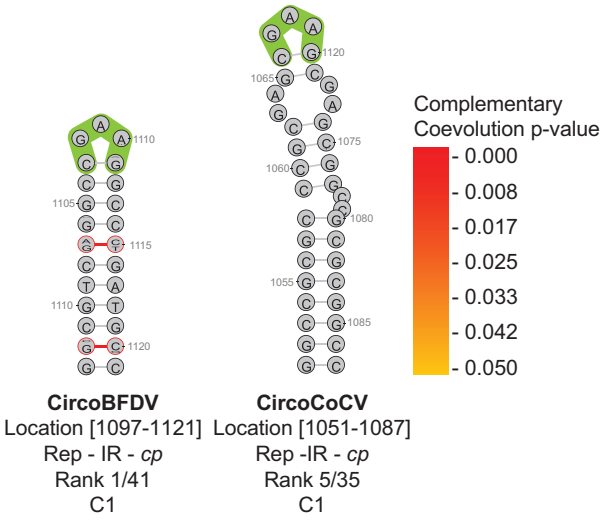
Stem loop
GeminiMYVYV T**AATAAA**GATTGAA**TTTTATT**T
GeminiTYLCV T**AATAAA**ATTTATAT**TTTTAT**
GeminiEACMV T**AATAAA**GGTTGAA**TTTTATT**G

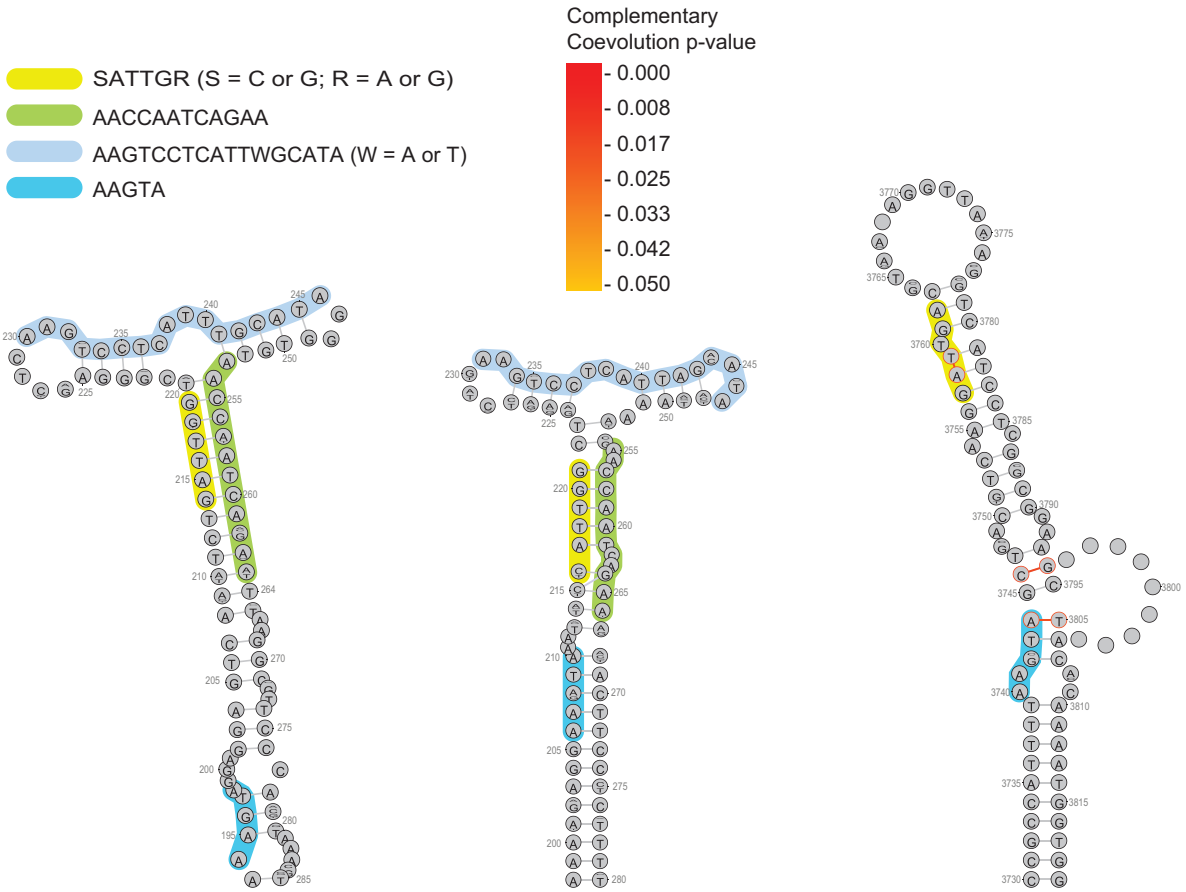
Stem loop
ATYTATTTCTATGAT (Y= C or T)

ATACAYTGT (Y=C or T)

Stop codon







AnelloTTSuV1
Location: [193-285]
Rank 7/20
A1

AnelloTTSuV2
Location: [198-280]
Rank 1/27
A1

AnelloTTV
Location: [3730-3819]
Rank 7/78
A1

Table 1. List of the 23 of large datasets obtained.

	Name[a]	Families	Constituent virus species	Size[b]
1	CircoPCV	Circoviridae	*Porcine circovirus 2*	519
2	CircoCoCV		*Columbid circovirus*	36
3	CircoDGCV		*Duck circovirus, Goose circovirus, Muscovy duck circovirus, Cygnus olor circovirus*	49
4	CircoBFDV		*Beak and feather disease virus*	184
5	AnelloTTSuV1	Anelloviridae	*Torque teno sus virus 1*	21
6	AnelloTTSuV2		*Torque teno sus virus 2, Porcine torque teno virus 2*	44
7	AnelloTTV		*Torque teno virus*	22
8	ParvoAAV	Parvoviridae	*Adeno-associated virus,*	34
9	ParvoHBoV		*Human boccavirus - 2, 3,4; Porcine boccavirus 1,2; Gorilla boccavirus, Bovine parvovirus 1, Canine minute virus*	21
10	ParvoMPV		*Mouse parvovirus 4, Rat minute virus, Mouse parvovirus, Minute virus, LuIII virus, Hamster parvovirus*	26
11	NanoBBTV-R	Nanoviridae	*Banana bunchy top virus component R*	221
12	NanoBBTV-S		*Banana bunchy top virus component S*	189
13	NanoBBTV-M		*Banana bunchy top virus component M*	150
14	NanoBBTV-N		*Banana bunchy top virus component N*	148
15	NanoBBTV-C		*Banana bunchy top virus component C*	122
16	GeminiMSV	Geminiviridae	*Maize streak virus*	759
17	GeminiWDV		*Wheat dwarf virus*	138
18	GeminiPanSV		*Panicum streak virus*	41
19	GeminiTYDV-CpCV		*Tobacco yellow dwarf virus, Chickpea chlorosis virus, Chickpea yellows virus*	41
20	GeminiCpCDV		*Chickpea chlorotic dwarf virus*	43
21	GeminiTYLCV		*Tomato yellow leaf curl virus*	228
22	GeminiEACMV		*East African cassava mosaic virus, South African cassava mosaic virus*	146
23	GeminiMYVYV		*Malvastrum yellow vein Yunnan virus, Cotton leaf curl Multan virus isolate, Bhendi yellow vein India virus*	254

[a] The name of the dataset is made up of the prefix of its family and the abbreviation of the main virus species it contains.

[b] The number of full genome sequences in the dataset

Table 2. Tajima's D – Fu and Li statistics for paired and unpaired genomic site alignments

		Tajima's D			Fu & Li's F		
	Large dataset	**Paired[a]**	**Permuted unpaired[b]**	***p*-value**	**Paired[c]**	**Permuted unpaired[d]**	***p*-value**
1	CircoPCV	-2.08	-1.90	0.06	-5.33	-4.91	0.13
2	CircoCoCV	-2.48	-1.75	< 0.01	-4.98	-3.17	< 0.01
3	CircoDGCV	1.44	0.71	0.99	1.40	0.58	0.96
4	CircoBFDV	-1.53	-1.33	< 0.01	-2.54	-1.77	< 0.01
5	AnelloTTSuV1	-0.72	-1.09	1	-0.19	-1.09	1
6	AnelloTTSuV2	-0.96	-0.95	0.54	-1.09	-1.38	0.92
7	AnelloTTV	-0.10	-0.09	0.5	-0.95	-0.95	0.49
8	ParvoAAV	-0.56	-0.55	0.53	0.80	0.67	0.85
9	ParvoHBoV	-1.09	-1.02	0.07	0.25	0.18	0.78
10	ParvoMPV	-0.22	-0.12	0.08	-0.31	-0.01	0.01
11	NanoBBTVR	-1.31	-1.21	0.28	-3.08	-2.27	0.03
12	NanoBBTVS	-0.96	-0.65	< 0.01	-2.79	-2.00	0.02
13	NanoBBTVM	-0.77	-0.38	0.05	-2.32	-1.74	0.14
14	NanoBBTVN	-1.61	-1.45	0.13	-4.26	-3.36	0.02
15	NanoBBTVC	-1.44	-0.80	< 0.01	-5.03	-3.28	< 0.01
16	GeminiMSV	-2.02	-1.72	< 0.01	-5.34	-3.26	< 0.01
17	GeminiWDV	-1.26	-0.61	< 0.01	-4.03	-2.99	0.04
18	GeminiPanSV	-0.91	-0.61	< 0.01	-0.53	-0.05	< 0.01
19	GeminiTYDV-CpCV	-1.28	-0.55	< 0.01	-2.30	-0.29	< 0.01
20	GeminiCpCDV	-0.88	-0.10	< 0.01	-0.71	0.19	< 0.01
21	GeminiTYLCV	-1.67	-1.71	0.61	-3.44	-3.23	0.26
22	GeminiEACMV	-1.33	-0.83	< 0.01	-2.58	-1.31	< 0.01
23	GeminiMYVYV	-1.23	-0.78	< 0.01	-3.25	-1.28	< 0.01

[a]Tajima's D for paired-sites alignments corresponding to the HCSS.

[b]Average Tajima's D for 100 permuted alignments sampled for the unpaired-sites.

[c]Fu and Li's F for paired-sites alignments corresponding to the HCSS.

[d]Average Fu and Li's F for 100 permuted alignments sampled for the unpaired-sites

Table 3. Comparison of synonymous substitution rates at paired and unpaired codon-sites.

	Datasets	Genes studied	Number of sequences	PARRIS[a]	FUBAR[b]
1	CircoPCV	*rep*, *cp*	30, 29	*rep*	*rep*
2	CircoCoCV	*rep*, *cp*	30, 30	*rep*	*rep*
3	CircoDGCV	*rep*, *cp*	30, 30	*rep*	*rep*
4	CircoBFDV	*rep*, *cp*	30, 29	*rep*, *cp*	*rep*, *cp*
5	AnelloTTSuV1	ORF1	17	ORF1	ORF1
6	AnelloTTSuV2	ORF1	30	ORF1	ORF1
7	AnelloTTV	ORF1	21	ORF1	ORF1
8	ParvoAAV	*ns1, vp1*	23, 30	*ns1, vp1*	*ns1, vp1*
9	ParvoHBoV	*ns1, np1*	21, 21	*ns1*	*ns1*
10	ParvoMPV	*ns1, vp2*	25, 18	*ns1*	*ns1, vp2*
11	NanoBBTV-R	*rep*	28	*rep*	*rep*
12	NanoBBTV-S	*cp*	29	*cp*	*cp*
13	NanoBBTV-M	*mp*	27	*mp*	*mp*
14	NanoBBTV-N	*nsp*	27	-	-
15	NanoBBTV-C	*clink*	30	-	-
16	GeminiMSV	*rep*, *cp*, *mp*	30, 30, 30	*rep*, *cp*	*rep*, *cp*, *mp*
17	GeminiWDV	*rep*, *cp*, *mp*	30, 30, 30	*cp*, *mp*	*cp*, *mp*
18	GeminiPanSV	*rep*, *cp*, *mp*	30, 30, 30	-	-
19	GeminiTYDV-CpCV	*rep*, *cp*, *mp*	30, 30, 30	*cp*	*cp*
20	GeminiCpCDV	*rep*, *cp*, *mp*	30, 30, 30	*rep*, *cp*, *mp*	*rep*, *cp*, *mp*
21	GeminiTYLCV	*rep*, *cp*	27, 30	-	-
22	GeminiEACMV	*rep*, *cp*	30, 30	*rep*	*rep*
23	GeminiMYVYV	*rep*, *cp*	29, 28	*rep*	*rep*

[a]Gene alignments in which the synonymous substitution rates (computed using PARRIS) at paired codon sites are significantly (Mann Whitney U test p-value < 0.05) lower than those at unpaired codon sites.

[b]Gene alignments in which the synonymous substitution rates (computed using FUBAR) at paired codon sites are significantly (Mann Whitney U test p-value < 0.05) lower than those at unpaired codon sites.

Table 4. Association between paired sites and complementarily coevolving sites.

	Dataset	Chi-squared value	*p*-values
1	CircoPCV	190.9307	4.20×10^{-14}
2	CircoCoCV	0.2272	0.14
3	CircoDGCV	143.2324	3.15×10^{-14}
4	CircoBFDV	62.5998	1.59×10^{-13}
5	ParvoAAV	185.5472	4.08×10^{-14}
6	ParvoHBoV	96.656	2.13×10^{-14}
7	ParvoMPV	137.077	3.02×10^{-14}
8	AnelloTTSuV1	117.9971	2.60×10^{-14}
9	AnelloTTSuV2	38.2243	2.41×10^{-08}
10	AnelloTTV	70.6212	1.55×10^{-14}
11	NanoBBTV-R	107.8986	2.37×10^{-14}
12	NanoBBTV-S	20.398	1.28×10^{-04}
13	NanoBBTV-M	49.9491	7.88×10^{-11}
14	NanoBBTV-N	48.2752	1.79×10^{-10}
15	NanoBBTV-C	21.1911	8.81×10^{-05}
16	GeminiMSV	212.2187	4.67×10^{-14}
17	GeminiWDV	89.9702	1.98×10^{-14}
18	GeminiPanSV	82.3437	1.81×10^{-14}
19	GeminiTYDV-CpCV	28.6975	2.43×10^{-06}
20	GeminiCpCDV	98.1122	2.16×10^{-14}
21	GeminiTYLCV	159.2665	3.50×10^{-14}
22	GeminiEACMV	175.6639	3.86×10^{-14}
23	GeminiMYVYV	364.9167	8.03×10^{-14}

Table 5. Summary of results.

	Dataset	>5 NASP structures[a]	dS paired codon sites< dS unpaired codon sites [b]	Selection at paired sites[c]	Complementary coevolution[d]
1	CircoPCV	+	+	-	+
2	CircoCoCV	+	+	+	-
3	CircoDGCV	-	+	-	+
4	CircoBFDV	+	+	+	+
5	AnelloTTSuV1	+	+	-	+
6	AnelloTTSuV2	+	+	-	+
7	AnelloTTV	+	+	-	+
8	ParvoAAV	+	+	-	+
9	ParvoHBoV	+	+	-	+
10	ParvoMPV	+	+	+	+
11	NanoBBTV-R	+	+	+	+
12	NanoBBTV-S	+	+	+	+
13	NanoBBTV-M	+	+	+	+
14	NanoBBTV-N	+	-	+	+
15	NanoBBTV-C	+	-	+	+
16	GeminiMSV	+	+	+	+
17	GeminiWDV	+	+	+	+
18	GeminiPanSV	+	-	+	+
19	GeminiTYDV-CpCV	+	+	+	+
20	GeminiCpCDV	+	+	+	+
21	GeminiTYLCV	+	-	-	+
22	GeminiEACMV	+	+	+	+
23	GeminiMYVYV	+	+	+	+

[a] datasets that had more than 5 structures significantly conserved in all lineages are given a "+" sign

[b] datasets in which at least for one gene alignment the synonymous substitution rates at paired codon sites were significantly lower than those at the unpaired codon sites are given a "+" sign

[c] datasets in which purifying selection detected within paired nucleotide sites was significantly stronger than that at unpaired nucleotide sites based on F and D statistics are given a "+" sign

[d] datasets in which a statistically significant association between paired sites and complementarily coevolving sites is detected are given a "+" sign